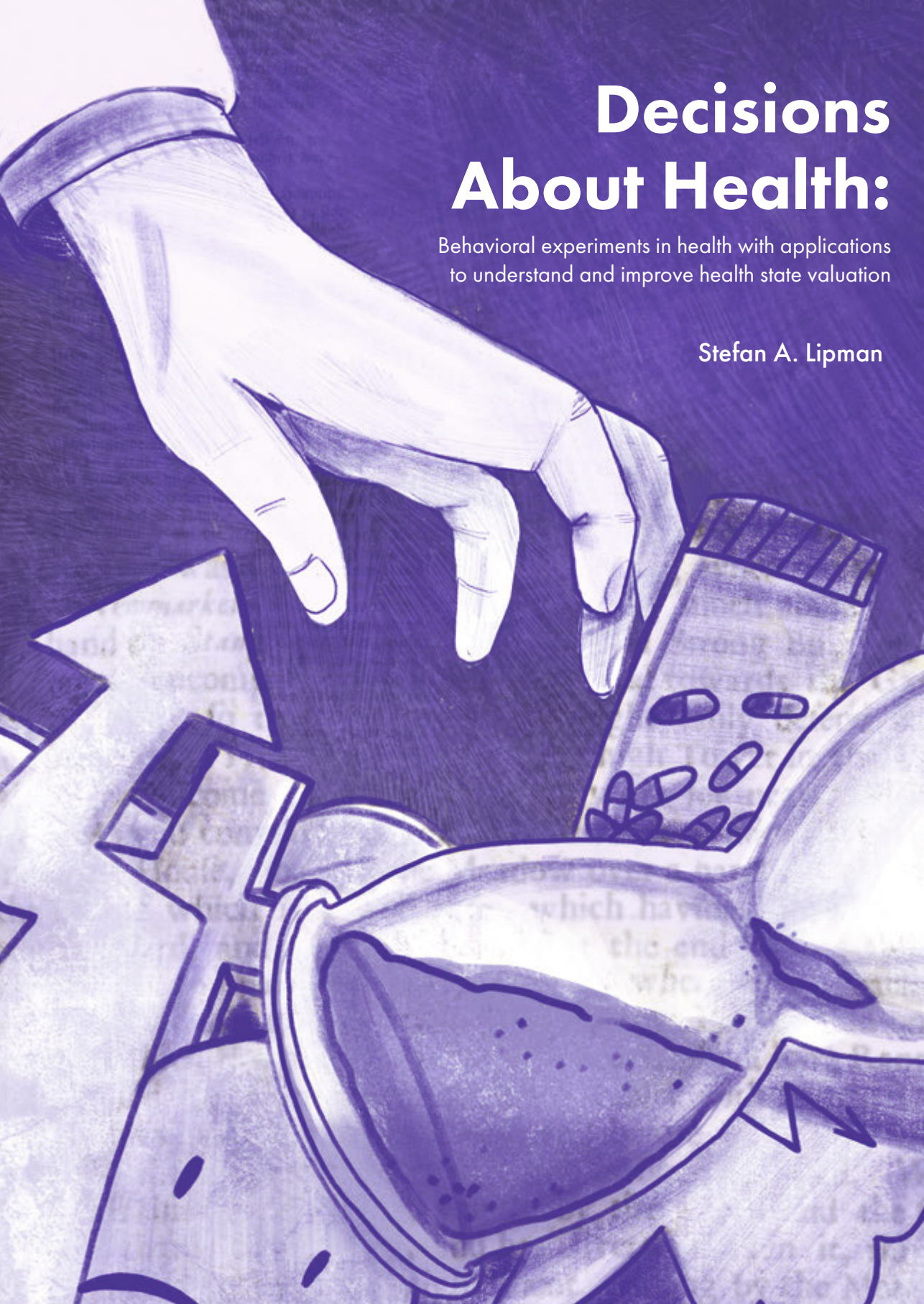


Decisions About Health:

Behavioral experiments in health with applications
to understand and improve health state valuation

Stefan A. Lipman



Decisions About Health:

Behavioral experiments in health with applications to understand and improve health state valuation

Stefan A. Lipman

COLOFON

ISBN: 978-94-6416-025-3

© Stefan A. Lipman

No part of this thesis may be reproduced or transmitted in any forms or means without permission of the author or the corresponding journal

Cover: © Anna Lena Illustrations

Printed by Ridderprint BV, Ridderkerk, The Netherlands

Decisions about Health
Behavioral experiments in health with applications to understand and improve health state
valuation

Keuzes over gezondheid
Gedragsexperimenten op het gebied van gezondheid met toepassingen om
gezondheidswaardering te begrijpen en te verbeteren

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof. dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
15 oktober 2020
om 15:30 uur

door

Stefan Adriaan Lipman
geboren te Gouda.

Promotiecommissie:

Promotor: prof. dr. W.B.F. Brouwer

Overige leden: prof. dr. N.J.A. Van Exel
prof. dr. J.H. Cawley
prof. dr. K.I.M. Rohde

Copromotor: dr. A.E. Attema

Contents

Part I: Behavioral experiments in health	22
Chapter 1: Introduction	7
Chapter 2: Rabin’s Paradox for Health Outcomes	22
Chapter 3: A QALY LOSS IS A QALY LOSS IS A QALY LOSS: A note on independence of loss aversion from health states	35
Chapter 4: One size fits all? Designing financial incentives tailored to individual preferences	47
Chapter 5: Trust me; I know what I am doing. Does domain experience reduce preference reversals in decision making for others?	59
Part II: Applying behavioral insights to health state valuation	77
Chapter 6: What’s it going to be, TTO or SG? A direct test of the validity of health state valuation	79
Chapter 7: QALYs without Bias? Non-parametric correction of time trade-off and standard gamble weights based on prospect theory	89
Chapter 8: The corrective approach: policy implications of recent developments in QALY measurement based on prospect theory.	105
Chapter 9: Living up to expectations: Experimental tests of subjective life expectancy as reference point in time trade-off and standard gamble	117
Chapter 10: A comparison of individual and collective decision making for standard gamble and time trade-off	143
Chapter 11: Discussion	159
Summary	173
Nederlandse samenvatting	177
Portfolio	183
Dankwoord	189
References	193
About the author	210



CHAPTER 1

Introduction

Every Monday, I start my day with a light breakfast (fat-free quark), grab a quick coffee, and take the bike to the train station in my hometown. After a short trip by train, and another trip on my bike, I'm ready to start a day at the office. Most time I spend sitting behind my desk, working on revisions or preparing for teaching. Even though I have a desk that can be adjusted to standing-mode, I prefer to work seated. A steady stream of coffee, water and (mostly healthy) snacks keep me energized, hydrated and satiated. After work, usually I can be found in the gym, to let off some steam resulting from the latest journal rejection, just before I head back home and prepare for volleyball practice.

You may ask, is there a point to this insight into the rather monotone life of an early career researcher? It stands to show that even a dusty academic's day is filled with decisions that are either about health or will have a direct effect on health. For example, I choose to work seated, while I know that being sedentary for a prolonged time may have detrimental health effects (Van Uffelen et al., 2010), and multiple times each week I'm faced with the dilemma to go to the gym in Rotterdam and improve my health, or head back home as quickly as possible.

My dissertation deals with such *decisions about health* and provides some insights from a behavioral and health economic perspective. My focus is both on individual decisions about health (e.g. 'Should I have surgery or ask for radiation therapy?' or 'Should I invest in my health today to reap the rewards later?') and on decisions about health at the societal level (e.g. 'Is this life-improving drug too expensive to fund from public health care resources?' or 'Which provides more health for society, surgery or radiation therapy?').

The main goal of this dissertation is extending and applying the methods and theories from behavioral economics to improve understanding of individual and societal decisions about health. In this introduction of my dissertation I will first attempt to convince the reader of why studying decisions about health is important and relevant, followed by a short characterization of how economists typically studied decision-making (about health). As an illustration, I will try to apply this approach to my decision to work seated rather than standing. Having established the main assumptions present in the traditional approach to studying decisions about health, I will discuss how these assumptions are also relevant for decision-making about health at a societal level (in the context of economic evaluations). The short summary of existing work in behavioral economics that follows, however, shows that these assumptions are often a poor description of how individuals *actually* decide about health. Instead, behavioral economics attempts to incorporate insights from psychology and other behavioral sciences (i.e. behavioral insights) into economic theory and methods to improve their applicability to actual decisions and behavior. The use of such behavioral insights in health economics is relatively novel, but appears to be a promising way forward. This dissertation (consisting of two distinct parts) aims to contribute to this growing field, and in particular to behavioral health economics.

Decisions about health – why should we study them?

The importance of studying decisions about health is best illustrated by two global trends: a) the increase in relative mortality (Bennett et al., 2018) as a result of preventable non-communicable diseases (such as cardiovascular disease, diabetes and cancer), and b) the rise in proportions of gross domestic products spent on health and health care (WHO, 2018). Whereas the former indicates that more and more individuals are becoming ill and/or dying

from causes that could have been prevented by changing health behavior, the latter requires societies to decide how to spend healthcare resources sensibly. To curb these trends, understanding how individuals decide about their health and how to promote effective societal decision-making about health could be crucial.

Furthermore, studies of health and decisions about health are of obvious importance, as being healthy is rated by many individuals as one of the most important goals in life (Bowling, 1995), is one of the most important contributors to wellbeing (Dolan et al., 2008), and is often a prerequisite to strive towards realizing other goals and psychological needs, such as being appreciated or self-actualization (Maslow, 1943). Indeed, the World Health Organization (1948) considers the highest attainable standard of health a fundamental right of every human being. Nonetheless, many individuals (myself included) often show preferences and behavior that is in stark contrast with the importance health has in individuals' lives and society. For example, only half of the Dutch population meets national exercise guidelines (RIVM, 2015). Furthermore, still 23% of Dutch people smoke (RIVM, 2015), and only 15% consume the recommended amount of vegetables or fruit on a daily basis (Van Rossum et al., 2017).

Rational decision making (about health)

Decisions about health (or with consequences for health), such as whether to smoke or not, to have a burger instead of fat-free quark, or to work sedentary or standing, usually involve costs and benefits both on the short and long term for the individual involved. For example, I am one of the lucky few to work at a desk that could be used while standing. Standing up from my office chair has the benefit of improving my health, but on the other hand standing requires more effort from me than sitting does (and the desk makes a terrible noise when adjusted to standing mode). In economic theory and applications thereof, any decision, including my decision whether or not to work standing, is seen as a reflection of the trade-off of these costs and benefits (Mankiw, 2020, Rubinstein, 2012). Traditionally, it was assumed that individuals engage in such trade-offs in a perfectly rational manner (Savage, 1954, von Neumann and Morgenstern, 1944), although economists have disagreed on what exactly such rational decision making entails (Wakker, 2010). Below I provide a characterization of such an individual, similar to those that were traditionally assumed to inhabit the theories and experiments developed by (health) economists. We will refer to this individual, as is often done (e.g. Thaler, 2000, Thaler and Sunstein, 2009) as *homo economicus*.

In order to formally model and predict decisions (about health) the following is often assumed about *homo economicus* (as summarized by Galizzi, 2014, Lazear, 2000):

- when faced with a set of options to choose from, *homo economicus* has complete, coherent, stable and consistent preferences;
- taking into account all information, *homo economicus*'s preferences maximize utility (or alternatively wellbeing), such that the final choices can be seen as the best possible option in terms of costs and benefits;
- when (as is usually the case with decisions for health) the available options are uncertain or risky (i.e. have multiple possible outcomes), *homo economicus* considers all outcomes and weights them by their likelihood of occurring. The option that yields the highest likelihood-weighted utility or wellbeing is preferred.

This characterization of homo economicus has a few implications for what is seen as a rational decision about health, relevant to this dissertation. First, rational decision-making implies consistency, such that if one prefers A over B, one will do so consistently, repeatedly and independent of context (e.g. time and place). For example, if I prefer a seated desk over a standing desk, I should have this preference today, tomorrow and every other day (all other things equal). Second, preferences are procedurally invariant, i.e. they are independent of how they are elicited. Simplified, this means, for example, that my answer on all of the following should be ‘seated desk’:

- i) which do you want to have in your office, a seated desk or a standing desk?,
- ii) for which would you be willing to pay more, a seated desk or a standing desk?,
- iii) which do you find is worth more fat-free quark, a seated desk or a standing desk?

Given that I prefer a seated desk over a standing desk, I should pick the former over the latter. Similarly, as both money and fat-free quark have positive value for me, I assign more value to the seated desk in both monetary and dairy terms. Third, for health decisions that involve uncertainty or risk, often a specific theoretical approach is used to predict and prescribe individuals’ choices: Expected Utility (EU) theory. As is the case with most economic theories, EU is usually defined and illustrated algebraically. Such formal descriptions of economic theories or methods are printed in Boxes throughout this Introduction for interested readers (e.g. see Box 1.1 for a definition of EU theory), but they can be skipped without loss of continuity.

Box 1.1. Expected Utility (EU) theory

Throughout all boxes, we will denote preferences as usual, i.e. \succ , \succsim , and \sim denote strict preference, weak preference and indifference, respectively. If we assume that preferences satisfy EU theory, gambles of the form $x_p y$, i.e. gambles yielding outcome x with probability p and outcome y otherwise, are evaluated as:

$$EU(x_p y) = p * U(x) + (1 - p) * U(y).$$

Here, $U(\cdot)$ is a real-valued, monotonically increasing utility function that assigns to each outcome a real number that represents how valuable that outcome is. As such, if we find an indifference between a gamble $x_p y$ and a sure outcome z , this implies:

$$EU(x_p y) = EU(z) \rightarrow p * U(x) + (1 - p) * U(y) = U(z).$$

Expected utility and QALYs

The implications of EU for rational decision-making about health are easily illustrated with a stylized example. Let’s assume for the sake of this example that my decision to work while seated or standing is going to have some influence on my health from 70 onwards (when I’m finally allowed to retire), and no influence on my health before age 70. If I choose to work while sitting for the rest of my career, I’m increasing my risk of cardiovascular disease. For now, assume that being sedentary will mean that on my 70th birthday, I have a 20% chance to have a fatal heart attack, and otherwise I will live out the remaining 20 years of my life in perfect health. If I choose to work while standing, I completely remove this risk of a fatal heart attack. However, the effort involved with working while standing is going to take a toll

on my knees, ankles and lower back, such that from my 70th birthday onwards I will need a cane to move about. For simplicity, we will abstract from all non-health outcomes¹, e.g. from the effort costs of working while standing, the annoying noise my standing desk makes, and assume that my preference for working while sitting is only related the health outcomes of seated and standing work.

Which of the two modes of working I will prefer, will depend on: i) how much I care about my health after 70, ii) how bad I feel it is to need a cane to walk (compared to perfect health), and iii) how much risk of dying I am willing to take. If, as my daily routine showed, I prefer working seated, this means that I assign more utility to living 20 years in perfect health with a 20% chance of a fatal heart attack (Option A) than living 20 more years whilst needing a cane (Option B). Or in other words, I find that option A is giving me more health utility than option B. But how much more, and how can we model such decisions in economic theory?

In health economics, such questions are answered by using quality-adjusted life years (QALYs) to express the amount of health received in some health profile. QALYs comprise both length and quality of life into a single measure. One finds the amount of QALYs associated with a health profile by multiplying the duration of the profile by a QALY weight, which represents the health-related quality of life experienced during that period. These QALY weights are normalized such that being dead has weight 0 and perfect health has weight 1 (with states worse than dead receiving negative weights). Under that normalization, 10 years in perfect health equals 10 QALYs. However, when health is less-than-perfect, this receives a weight smaller than 1, such that each year in that health state is worth less than 1 QALY. For example, if we assume that a disease reduces quality of life with 50% (compared to perfect health), each year someone lives in that health state is worth $\frac{1}{2}$ QALY (as the QALY weight for that health state is 0.5).

This simple and intuitive way of expressing health benefits is derived from the linear QALY model (see Box 1.2 for definition and example), which is a model of individual preferences that assumes that EU holds (Pliskin et al., 1980) and implies the following for decisions about health. First, the use of EU assumes that individuals have consistent preferences and are perfectly capable of probability calculus (i.e. are homo economicus). Second, the use of the linear QALY model typically implies that each year has the same value as the next, i.e. utility of life duration is linear. Third, a linear utility function implies risk neutrality for life duration (i.e. no risk seeking or risk aversion for years of life).

At this point, one may question why a model of individual preferences with such strict assumptions about how we decide about health is relevant for this dissertation. The straightforward answer to this question is that these assumptions allow simple measurement and use of QALYs in practice, as I will elaborate on below.

¹ Note that this simplification is in no way necessary to apply EU to study decisions about health and thus only reflects a narrative choice.

Box 1.2. linear QALY model

QALY models are usually applied to chronic health outcomes, i.e. outcomes that involve the experience of a single health state for some prolonged duration (as opposed to health outcomes characterized by short episodes of ill health, such as epilepsy). In the linear QALY model, we denote health outcomes as (T, Q) , i.e. T years in health state Q , and preferences for health outcomes can be represented by:

$$U(T, Q) = T * V(Q).$$

Here, $U(\cdot)$ and $V(\cdot)$ are the utility functions over health outcomes and health status respectively. Lotteries of the form $(T_1, Q_1)_p(T_2, Q_2)$, i.e. (T_1, Q_1) with probability p and (T_2, Q_2) otherwise (i.e. with probability $1 - p$) are evaluated by EU, i.e.:

$$p * U(T_1, Q_1) + (1 - p) * U(T_2, Q_2) = p * T_1 * V(Q_1) + (1 - p) * T_2 * V(Q_2).$$

The following normalization is often used with the linear QALY model: $U(\text{death}) = 0$ and $V(\text{perfect health}) = 1$. Applied to the example reported in text, we find:

$$(20, \text{perfect health})_{0.8}(\text{death}) > (20, \text{cane}),$$

which is evaluated by:

$$0.8 * 20 * V(\text{perfect health}) + (1 - 0.8) * V(\text{death}) > 20 * V(\text{cane}).$$

We can simplify this to:

$$16 > 20 * V(\text{cane}), \text{ i.e. } V(\text{cane}) < 0.8.$$

Societal decisions about health: cost-utility analyses

The growing pressure on public spending on health care necessitates considering if new and existing treatments provide sufficient value for money. Such considerations are complex, as they require both an accurate assessment of the costs and benefits associated with a treatment, and a means of determining when value for money is 'sufficient'. For example, consider a recent innovation in the treatment of spinal muscular atrophy in infants, which was priced at almost 2 million euro per patient (Cohen, 2019). This one-time treatment can result in drastic improvements in otherwise soon paralyzed and terminally ill infants, and hence appears to provide substantial value. However, funding this treatment from public health care resources would considerably affect the available budget (which could also be used for treating other patients or hire more staff in elderly homes). Hence, the question whether the value provided is sufficient to warrant reimbursement becomes crucial in these societal decisions about health.

Economic evaluations, in which a comparison is made between the costs associated with a treatment and treatment-related benefits, are increasingly often used in this context (Drummond et al., 2015). Treatment-related benefits can be expressed in different ways, but for the sake of comparability across treatments and patient groups the QALY is often chosen as outcome measure. Economic evaluations that inform policymakers about the incremental costs per QALY gained by some treatment (compared to a relevant comparator), are often referred to as cost-utility analyses (CUAs). These incremental costs per QALY are often compared against a threshold for reimbursement, which may differ between countries

(Drummond et al., 2015). Reimbursement decisions, furthermore, may also depend on other factors the public may believe to be relevant: e.g. the age of the recipient of the treatment, how severe the consequences are of not reimbursing treatment, and how rare the condition being treated is (van de Wetering et al., 2016).

To illustrate how CUAs inform societal decisions about health and discuss several methodological issues related to this dissertation, let us consider again the standing desk. For example, imagine that the Dutch Ministry of Health is considering the health benefits of providing all public offices with standing desks to reduce prolonged sedentariness (disregarding that like mine they may remain unused). In order to determine the cost-effectiveness of this policy using CUA, the health benefits of having a standing desk need to be expressed in QALYs. Again, let us assume that this benefit is captured in a reduced risk of cardiovascular disease at age 70, but at a loss of mobility (i.e. needing a cane to walk about). Calculating the QALYs associated with having (and using) a standing desk then requires determining the utility associated with needing a cane.

A difficult question, however, is whose utility should matter (Versteegh and Brouwer, 2016). First, one could consider to take into account the utility of the patients who would benefit from treatment (Aronsson et al., 2015, Leidl and Reitmeir, 2011). For example, if we want to perform a cost-utility analysis of the health benefits of standing desks under the assumptions discussed above, this requires assessing the utility associated with needing a cane to walk about (from age 70 and onwards). Hence, in CUAs one could take into consideration the utility that individuals who need a cane to walk about assign to their health status. However, an often-voiced concern is that patients adapt to their health status (Damschroder et al., 2005, Damschroder et al., 2008, Menzel et al., 2002). For example, a bed-ridden patient may be, altogether, quite happy (see the classic work comparing paralyzed accident victims and lottery winners: Brickman et al., 1978), which is a remarkable testament of humans' ability to strive in tough situations. As such, by only including patients in attempts to measure health utilities (i.e. health state valuation) we may assess the possible benefits accrued by improving their condition as being relatively low. For example, if I need a cane for the last 20 years of my life, this might have a strong negative effect on my utility at first, but over time I may get used to needing the cane and find ways of dealing with this such that it impacts my health utility less. If asked to value my health, at this later point in time, I might provide valuations as high as feeling perfectly healthy. If a treatment becomes available that allows me to walk again, it may appear that relatively little is gained in terms of utility compared to the initially already high valuation of needing a cane.

Since no consensus exists on if and how to correct for adaptation (Versteegh and Brouwer, 2016), instead QALYs are often derived from preferences of the general public. Given that health-care costs are provided for by all members of the general public collectively, it could be argued that their utilities should be used to reflect the preferred societal perspective in CUAs. For example, if the Dutch Ministry of Health should provide standing desks in all public buildings, the utility the general public (that would collectively finance the desks) assigns to the standing desks' health benefits could be taken into account. Typically, such QALY weights for the general public are obtained using a representative sample that values hypothetical health states by imagining themselves living in these states for some duration (Oppe et al., 2014, Stolk et al., 2019).

Health state valuation: time trade-off and standard gamble

Two of the most popular methods used for health state valuation are time trade-off (TTO) and standard gamble (SG)². In the TTO method, respondents are asked to imagine living in impaired health for some fixed duration. Alternatively, respondents can choose to live for a shorter period in perfect health. This shorter time in full health is varied until respondents indicate that they consider both health profiles equivalent. For example, imagine you would live for 10 more years, and you would need a cane to walk about. However, you are offered a treatment that will allow you to walk without a cane, but if you take this treatment you will live for a shorter amount of time. How many years in perfect health do you consider to be equivalent to living 10 more years with a cane? Perhaps you found 7 years in perfect health equivalent to 10 years whilst needing a cane. This implies that your QALY weight for requiring a cane to walk is $7/10 = 0.7$ (see Box 1.3 for the justification for this derivation).

In the SG method, again, respondents are asked to imagine living in some impaired health state for a fixed duration. Unlike in the TTO method, for SG they are now offered an alternative treatment, which is risky. This treatment takes the form of a gamble, which either yields perfect health for the same fixed duration with some probability or immediate death otherwise. The risk involved with this gamble is varied until respondents indicate they find both options to be equivalent. For example, again imagine living 10 years whilst needing a cane. Would you undergo a treatment to cure your mobility problems (for your remaining 10 years of life), if the chance exists you might die immediately instead (e.g. as a result of undergoing risky surgery)? And if so, what is the largest chance of immediate death you would be willing to risk? Perhaps you were only willing to take the gamble (to cure your mobility problems) if the treatment has a probability of success larger than 85% (i.e. a risk of death of less than 15%). This would imply your QALY weight for life with a cane is 0.85 (see Box 1.3 for justification).

Although TTO and SG share their purpose, i.e. eliciting QALY weights, the methods typically yield different results when applied to the same health state. As you might have experienced when considering the examples above, often SG yields higher QALY weights compared to TTO (e.g. Bleichrodt and Johannesson, 1997, Torrance, 1976). This poses a problem to those relying on CUAs to inform policy, as health benefits expressed in QALYs become dependent on the method chosen to value them. As such, most institutions that perform and evaluate CUAs (such as NICE in the UK and the Healthcare Institute in the Netherlands) pick a single instrument to measure and express health benefits with. Examples of such instruments are the EQ-5D and the Short Form Six Dimensions (SF-6D), which capture relevant facets of health-related quality of life along a few clearly defined dimensions. Such instruments are generic, meaning that they can be applied to describe health-related quality of life for a multitude of diseases. The QALY weights for such generic measures are often obtained “indirectly” from nationally representative tariff lists, which are obtained in studies conducted in samples of the general adult population (Stolk et al., 2019). However, given that different methods are used for generating these tariff lists (e.g. EQ-5D involves TTO, and SF-6D uses SG), and no general consensus exists to prefer one over the other, this is only a partial solution. In my dissertation I explore alternative solutions to these

² Pictorial representations of these methods can also be found in the Online Supplements of this dissertation.

methodological problems, by exploring the role of the perhaps unrealistic assumptions used in the linear QALY model (described in Box 1.2).

Box 1.3. Deriving QALY weights using time trade-off and standard gamble

The TTO method, asks for a time equivalent in perfect health which yields indifference between (T_1, Q) and $(T_2, \text{perfect health})$, with $T_1 > T_2$. The number of years T_2 is varied until the respondent is indifferent (denoted by \sim). Under the linear QALY model, we evaluate TTO indifferences of the form $(T_1, Q) \sim (T_2, \text{perfect health})$ as follows:

$$T_1 * V(Q) = T_2 * V(\text{perfect health}).$$

This allows deriving the utility of health state Q as: $V(Q) = T_2/T_1$. Applied to the example reported in text, we find $(10, \text{cane}) \sim (7, \text{perfect health})$, which yields: $10 * V(Q) = 7 * V(\text{perfect health})$. Assuming $V(\text{perfect health}) = 1$, this simplifies to $V(\text{cane}) = 7/10$.

The SG method involves determining probability p at which decision makers are indifferent between a sure outcome (T_1, Q) , and a risky prospect $(T_1, \text{perfect health})_p(D)$. Under the linear QALY model, we evaluate SG indifferences of the form $(T_1, Q) \sim (T_1, \text{perfect health})_p(D)$ by:

$$T_1 * V(Q) = p * T_1 * V(\text{perfect health}) + (1 - p) * V(\text{death}).$$

This allows deriving the utility of health state Q as: $V(Q) = p$. Applied to the example reported in text, we find $(10, \text{cane}) \sim (10, \text{perfect health})_{0.85}(D)$, which yields: $10 * V(Q) = 0.85 * 10 * V(\text{perfect health})$. Assuming $V(\text{perfect health}) = 1$ and $V(\text{death}) = 0$, this simplifies to $V(\text{cane}) = 0.8$.

Beyond homo economicus for decisions about health?

Having briefly introduced the traditional economic approach to studying decisions about health, the overall goal of this dissertation can be formulated, i.e. extending and applying the methods and theories from behavioral economics to improve (understanding of) individual and societal decisions about health. Achieving this goal requires moving beyond the assumptions often used to study decisions about health, or in other words beyond homo economicus. This may be considered important and timely, as a plethora of evidence exists documenting violations of the assumptions that characterize homo economicus, which suggests that assuming individuals decide rationally about their health misrepresents real decisions. Below I provide a short review of a few of those findings relevant for this dissertation (more details can be found in the related chapters of my dissertation).

The following has been found for decision-making (about health):

- *We are risk averse for life duration*, i.e. we tend to avoid risks for life duration if we have the possibility. As is the case for monetary outcomes (Andersen et al., 2006, Andersen et al., 2008, Harrison et al., 2005), we often observe such risk aversion for life duration (Breyer and Fuchs, 1982, Kemel and Paraschiv, 2018, Oliver, 2018, van der Pol and Ruggeri, 2008, Verhoef et al., 1994). This means, for example, that when offered a treatment that will increase our life by 9 months, or a treatment that will yield 20 months with 50% chance and 0 months otherwise, most of us opt for the

certain treatment (even though the gamble is expected to yield more health). Such risk aversion violates the assumption of linear utility of life duration in the linear QALY model. Chapter 2 of this dissertation elaborates and provides additional evidence.

- *Procedural invariance is often violated*, i.e. we find preference reversals when comparing preferences between different methods. For example, Lichtenstein and Slovic (1971) found that respondents prefer a gamble with high certainty, but, when asked, provide higher monetary values to risky lotteries. This leads to a preference reversal, as preferred lotteries should also be valued higher (i.e. if you prefer A over B, you should also value A higher than B and vice versa). Their work inspired many replications and extensions for monetary outcomes (for a review, see: Seidl, 2002), and some studies have shown that such preference reversals also occur for decisions about health (Attema and Brouwer, 2013, Oliver, 2006, Oliver, 2013b). Chapter 5 of this dissertation discusses this further and provides additional evidence.
- *Reference-points matter*, i.e. health profiles are not necessarily evaluated in absolute terms (as in the linear QALY model) but rather in relative terms. Several studies have shown that health is often considered relative to a reference-point, i.e. a specific health outcome to which all other outcomes are compared. A few reference-points that have been suggested to affect decisions about health are: expectations about length (van Nooten and Brouwer, 2004, van Nooten et al., 2009) and quality of life (Wouters et al., 2015), the best (e.g. perfect health in TTO) or worst (e.g. immediate death in SG) possible outcome (van Osch et al., 2006), and the health individuals feel they deserve (Wouters, 2016). Chapters 3, 7, and 9 of this dissertation discuss this further and provide additional evidence.
- *Losses loom larger than gains*, i.e. it matters whether health profiles occur above or below the reference-point. Outcomes that exceed the reference-point are perceived as gains, while those that fall short of the reference-point are seen as losses. Many studies using monetary outcomes have shown loss aversion, i.e. many individuals are more sensitive to losses compared to similarly sized gains (Abdellaoui et al., 2008, Abdellaoui et al., 2016, Kahneman and Tversky, 1979, Tversky and Kahneman, 1992). Recently, this study of loss aversion has also been extended to health outcomes (Attema et al., 2013, Attema et al., 2016). Chapters 3 and 7 of this dissertation discuss this further and provide additional evidence.
- *Small changes in probabilities carry disproportionate weight*, i.e. small chances of good or bad health outcomes might be weighted heavily in decisions about health. Utility models based on EU (e.g. linear or generalized QALY models) do not allow for such probability weighting. This is likely to misrepresent decision-making as many studies have shown that individuals are especially sensitive to changes from impossible to possible (e.g. 0% to 1%) and uncertain to certain (e.g. 99% to 100%). This is referred to as inverse S-shaped probability weighting, which is the modal finding for decisions about money (Abdellaoui, 2000, Bleichrodt, 2001, Gonzalez and Wu, 1999, Suter et al., 2016) and decisions about health (Bleichrodt and Pinto, 2000, Bleichrodt et al., 1999, Suter et al., 2016). Chapters 5 and 7 of this dissertation discuss this further and provide additional evidence.

Many of these insights derived from experimental work in behavioral economics (i.e. behavioral insights) can be combined into a single alternative utility model that is central to

many of the chapters reported in this dissertation: prospect theory³. Several authors have suggested that prospect theory (or more accurately the behavioral insights captured in this model) could provide an explanation for several health-related decisions for which traditional economic theory has no explanation: e.g. low uptake of voluntary deductibles in health insurance (van Winssen et al., 2016), the too low or too high uptake of some forms of insurances (Gottlieb, 2012), the low uptake of screening (Baillon et al., 2018), and the difference between health state valuations derived with TTO and SG (Bleichrodt, 2002). At the time of writing of this dissertation, however, only a few studies have investigated the relevance of prospect theory to understand decisions about health empirically (Attema et al., 2013, Attema et al., 2016, Kemel and Paraschiv, 2018), as opposed to the larger evidence-base for monetary outcomes. As such, it is not entirely clear if the behavioral insights captured in prospect theory extend fully to health outcomes, and more behavioral experiments in health are thus needed (Galizzi and Wiesen, 2018).

Research objectives

With this thesis I therefore aim to provide i) additional understanding of how individuals **actually** decide about health (using theories and methods from behavioral economics), and ii) use this understanding to improve the methods used for health state valuations. These two research objectives are reflected in the structure of this dissertation, which consists of two parts.

In Part I, a series of ‘*Behavioral experiments in health*’ on varying topics is reported, while Part II has a specific focus on ‘*Applications of behavioral insights to health state valuation*’. Hence, whereas the first part of this dissertation presents research findings that may be relevant to understanding or improving decisions about health in many different contexts (e.g. decisions about length and quality of life, physician decision-making, and exercise behavior), the second part of this dissertation applies insights (derived in part from the studies reported in Part I) to the highly specialized context of health state valuation.

Part I of this dissertation reports on a series of behavioral experiments in health that provide some answers to the following research questions:

- To what extent are individuals risk averse for uncertain health outcomes with small or moderate stakes, and can EU explain such risk aversion? (Chapter 2)
- Does the degree to which individuals are loss averse for life duration depend on the quality of life experienced during this time? (Chapter 3)
- How heterogeneous are risk and time preferences, and can this heterogeneity be used to tailor financial incentives to improve decisions about health? (Chapter 4)
- Are decisions about health as (in)consistent as those for money, and does the degree of preference reversals depend on who makes these decisions? (Chapter 5)

Part II of this dissertation reports on a series of studies aimed at studying how behavioral insights could be utilized to obtain QALY weights that better reflect individuals’ trade-offs between length and quality of life, with the following research questions:

³ A full formal definition of prospect theory would go beyond the scope of this introduction. Chapters 3, 7 and 9 of this dissertation provide a formal application of prospect theory as a generalization of the (linear) QALY model for interested readers.

- Which method better reflects QALY weights according to respondents themselves: TTO or SG (Chapter 6)?
- Can prospect theory be applied to derive QALY weights with improved validity (i.e. corrected weights, Chapter 7)?
- How feasible is it to apply such ‘corrected weights’ in practice (Chapter 8)?
- What is the influence of expectations regarding length of life on TTO and SG weights (Chapter 9)?
- Are QALY weights improved by deciding on them collectively (Chapter 10)?

Outline of this dissertation

Part I (*Behavioral experiments in health*) starts with the extension to the health domain of a now-classic critique of EU theory, the concavity-calibration paradox presented by Rabin (2000), which provides a compelling case *against* EU and *for* reference-dependent theories (**Chapter 2**). **Chapter 3** continues this inquiry into the difference between decisions for health gains and health losses. This chapter reports a study that tested the stability of loss aversion, by exploring if the degree of loss aversion depends on the quality of life in which these lost years of life are spent. **Chapter 4** explores whether it is possible to design financial incentives that are tailored to individual preferences. Finally, Part I ends with a study on the degree of preference reversals in medical and financial decision-making for others, which is reported in **Chapter 5**.

Part II (*Applications of behavioral insights to health state valuation*) deals with health state valuation using TTO and SG. **Chapter 6** studies the degree to which QALY weights measured through these methods correspond to how individuals would value the corresponding health states on the 0-1 scale. **Chapter 7** reports the results of an experiment in which biases in TTO and SG weights are corrected using a model based on prospect theory. The findings of this study suggest that applying a ‘corrective approach’ to health state valuation, in which biases in measurement are approximated and corrected for based on prospect theory, may be a promising tool for improving economic evaluation. **Chapter 8** discusses this promise by showing how the decision to correct or not to correct influences outcomes of economic evaluations and discusses the many methodological and practical steps required before a corrective approach can be used to inform policy. One of these steps is addressed in **Chapter 9**, which studies whether expectations about length of life, i.e. one’s subjective life expectancy, serves as reference-point in health state valuations. Finally, in **Chapter 10** a different technique for resolving the difference between TTO and SG is studied. Instead of completing these health state valuation exercises by themselves, participants in this study completed TTO and SG in dyads, deliberating and bargaining about their decisions.

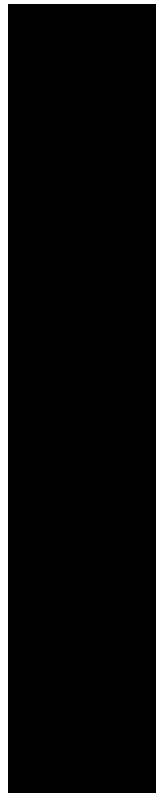
In the Discussion of this dissertation (**Chapter 11**), Part I and Part II are brought together again. After reflecting on the implications and limitations of my research regarding the overarching theme (understanding individual and societal decisions about health), this chapter concludes this dissertation.

Those of you who made it this far through my Introduction will be happy to hear there will be no more mentioning of the adjustable desk (which I am leaving as we speak, today is Monday, so I’m off to the gym)!



PART I

Behavioral experiments in health



2

Rabin's Paradox for Health Outcomes

Chapter based on:

Lipman, S.A., & Attema, A.E. (2019) Rabin's paradox for health outcomes.

Health Economics, 28(8), 1064-1071.

Abstract: Many health economic studies assume expected utility (EU) maximization, with typically a concave utility function to capture risk aversion. Given these assumptions, Rabin's paradox (RP) involves preferences over mixed gambles yielding moderate outcomes, where turning down such gambles imply absurd levels of risk aversion. Although RP is considered a classic critique of EU, no paper has as of yet fully tested its preferences within individuals. In an experiment we report a direct test of RP in the health domain, which was previously only considered in the economic literature, showing it may have pervasive implications here too. Our paper supports the shift towards alternative, empirically valid models, such as prospect theory, also in the health domain. These alternative models are able to accommodate Rabin's paradox by allowing reference-dependence and loss aversion.

Introduction

Risk is central in health economics, and is for instance covered in literature on health insurance (e.g. Arrow, 1963, Kairies-Schwarz et al., 2017), health state valuations (e.g. Pliskin et al., 1980, Torrance, 1976), health-related behavior change (e.g. Anderson and Mellor, 2008), and patient preferences (e.g. Galizzi et al., 2016b, Seston et al., 2007). Generally, individuals dislike risk, i.e. variance in outcomes, and prefer certain options with the same expected value over risky options. This is referred to as risk aversion, a hallmark phenomenon in economics that has found its way into many health applications, such as Arrow's (1963) classic exposition on health insurance.

Risk aversion is typically modeled within the framework of expected utility (EU) theory with a concave utility function over wealth (von Neumann and Morgenstern, 1944), often considered to be the normative and rational benchmark for decision making under risk (e.g. Harsanyi, 1955, Kahneman and Tversky, 1979, Wakker, 2010). However, the descriptive validity of EU, i.e. its applicability to understand or describe how individuals actually decide, has been questioned for decades (for a review, see: Starmer, 2000). For example, several paradoxes have been presented in the economic literature that violated EU, such as Allais' (1953) paradox and Rabin's paradox (RP, Rabin, 2000, Rabin and Thaler, 2001). As an illustration of the latter: consider an agent who turns down a 50/50 gamble of gaining 11\$ or losing 10\$ at all wealth levels. Rabin (2000) showed how under EU with concave utility this agent should turn any gamble with a 50/50 loss of 100\$, even when the agent could gain millions!

This thought experiment has become a classic criticism of EU as a descriptive theory of risk aversion. Although EU can capture risk aversion by assuming a concave utility function over wealth, this utility function should be extremely concave to capture turning down 50/50 gambles over such small stakes (e.g. gaining 11\$ or losing 10\$), which leads to absurd predictions for larger stakes. Given that most classic work on risk aversion in economics assumed EU with concave utility, RP has generated much debate among economists. Whereas some authors argue that EU is simply not a plausible theory for risk aversion and propose a move towards reference-dependent theories (e.g. Bleichrodt et al., 2019), others criticize the assumptions relevant to RP (Andersen et al., 2011, Harrison et al., 2017). However, empirical evidence on the presence of RP is scarce, only a handful of empirical studies are available that tested RP, all in the economic domain. First, Cox and colleagues (2013) observed RP preferences for financial outcomes in an incentive-compatible study. Second, Bleichrodt and colleagues (2019) identified the causes of RP empirically, showing how a reference-dependent model with loss aversion may explain RP (as already suggested by Rabin, 2000). A drawback of the first study, however, is that it involved highly unlikely outcomes (i.e. casino outcomes), while a drawback of the second study is that preferences for large stakes are not studied.

So far, both theoretical and empirical work has focused solely on Rabin's (2000, 2001) critique on EU in the monetary domain. There it has had vast implications, as citation scores show. It is well-known that preferences for health outcomes may differ from decisions for financial outcomes (even at a neurological level, see Suter et al., 2015). Such differences have been observed for time preferences (e.g. Attema et al., 2018b, Chapman, 1996), ambiguity preferences (e.g. Curley et al., 1984), and, most relevant to our purposes, risk

preferences (e.g. Suter et al., 2016, Weber et al., 2002). Whereas Allais' paradox has been tested with health outcomes (Oliver, 2003b), no such work exists for RP. Therefore, in this paper we extend RP to the health domain.

Notation and formal definitions

We first introduce notation and define RP for monetary outcomes (x, y) . We consider agents as modeled in EU (von Neumann and Morgenstern, 1944), who face gambles of the form $x_{0.5}y$, i.e. x results with probability 0.5 and y otherwise. Preferences are denoted by the usual \succ , \succsim , and \sim , representing strict preference, weak preference and indifference, respectively. Under EU, gambles are evaluated linearly in probabilities, i.e. $x_{0.5}y$ is evaluated by: $0.5 U(x) + 0.5 U(y)$, where often $U(\cdot)$ is assumed to be a strictly increasing and concave utility function over final wealth (which is equal to initial wealth I_w with the outcome of the gamble incorporated). This concavity of $U(\cdot)$ is assumed to reflect risk aversion, which is defined as preferring a gamble's expected value with certainty over the actual gamble (Wakker, 2010).

The classic RP thought-experiment starts with the assumption that an agent turns down a gamble $x_{0.5}y$ at all levels of initial wealth, i.e. always prefers staying at I_w over the gamble for some x and y . For example, assume $x = +11$ and $y = -10$, and consider someone who always turns down $(+11\$_{0.5} - 10\$)$. By means of a calibration process, Rabin (2000) showed that this person should also turn down gambles with extremely large expected value. To illustrate this calibration process, assume we observe such risk aversion for all I_w and utility over final wealth remains concave. Turning down $(+11\$_{0.5} - 10\$)$ at all wealth levels implies that over each length of 21 dollars U' drops by a factor of 10/11 (see Wakker, 2010). Such geometric decay is highly unlikely, as it implies that the marginal utility of each additional dollar diminishes expeditiously: take for example the decay of U' on an interval of 4200\$, which will be $\left(\frac{10}{11}\right)^{200} = 5.26 * 10^{-9}$ (more details in the Online Supplements of this dissertation). However, these conclusions only hold if gambles such as $+11\$_{0.5} - 10\$$ are indeed turned down at all wealth levels (or at least at many wealth levels⁴). Often, this empirical assumption is justified by observing that many agents will (ceteris paribus) reject such gambles at many (if not all) wealth levels, which led Rabin (2000) to assume that this gamble will also be turned down by a single agent at many (if not all) wealth levels.

Now, we extend RP to health outcomes (\mathcal{H}), which are quantifiable and real-valued (e.g. hours of life). We consider agents as modeled by EU in two cases: a) individual decisions – i.e. agents deciding about their own health, and b) societal decisions – i.e. agents deciding as societal decision makers for population health⁵. In both cases, $U(\mathcal{H})$ is a strictly increasing and concave utility function over final *health*. For individual decisions, initial health I_h

⁴ Much of the discussion surrounding RP has focused on this assumption, with its validity being questioned for example by: Andersen and colleagues (2011) and Harrison and colleagues (2017). Rabin (2000) showed that gambles need not be turned down at all wealth levels, and Wakker (2010) discusses how gambles only need to be turned down over relatively small domains of initial wealth to produce absurd concavity under EU.

⁵ See the Online Supplements of this dissertation for more detail on the social welfare function.

denotes an agent's life expectancy before a choice is considered, whilst for societal decisions I_h denotes the societal decision maker's judgement of society's initial health. In both cases final health is obtained by adding to I_h (gains) or subtracting from I_h (losses) the relevant health outcomes in gambles, i.e. $\mathcal{L}, \ell, I_h, \mathcal{G}, \mathcal{G} \in \mathcal{H}$ (see Table 2.1 for details on outcomes). We let \mathcal{G} (\mathcal{G}) represent a moderate (large) health gain compared to initial health I_h and we let ℓ (\mathcal{L}) denote a moderate (large) health loss compared to I_h . As in the canonical example by Rabin and Thaler (2001), we test RP by setting $\mathcal{G} = +11$ and $\ell = -10$ (e.g. +11 or -10 hours of life). Like Rabin (2000) for monetary outcomes, we assume that if $\mathcal{G}_{0.5}\ell$ is turned down by agents with many different levels of I_h , this implies that such gambles are also turned down by one individual at many life expectancies (for individual decision) and for many society's initial health levels (for societal outcomes)⁶. Under these assumptions (according to Rabin's (2000) calibration theorem), if we replace gamble $\mathcal{G}_{0.5}\ell$ with $\mathcal{G}_{0.5}\mathcal{L}$, with $\mathcal{L} = -100$, this person should turn down gambles for any \mathcal{G} (up to $\mathcal{G} = \infty$). Given the difficulties with grasping infinity, we elicit RP with $\mathcal{G} = 10,000$.

We define RP as the following combination of preferences: $I_h > \mathcal{G}_{0.5}\ell$ and $I_h < \mathcal{G}_{0.5}\mathcal{L}$, which constitutes a violation of EU with concave utility⁷. Whenever subjects turn down (accept) both gambles (i.e. $I_h > (<) \mathcal{G}_{0.5}\ell$ & $I_h > (<) \mathcal{G}_{0.5}\mathcal{L}$), we will say that they do not violate EU.

Method

Sample: $N = 201$ subjects were recruited by means of the Erasmus Research Participation System. All subjects were Business Administration students and were rewarded course credits for participation. The mean age of our sample was 20.29 (SD = 1.36) and 34% of our sample was female.

Procedure and Design: This experiment was part of a larger study on preferences for health outcomes, and was completed using Qualtrics Survey Software. Each subject completed all 6 RP gamble-pairs, which each consisted of a moderate stake gamble ($\mathcal{G}_{0.5}\ell$) and a calibrated large stake gamble ($\mathcal{G}_{0.5}\mathcal{L}$). The RP gamble-pairs were grouped in two counter-balanced blocks (completed within-subjects): 3 individual gamble-pairs and 3 societal gamble-pairs (presented in random order).

⁶ Obviously, for health outcomes there is less to no evidence that such preferences hold for many individuals at many levels of initial health. In fact, some authors have suggested that utility might be kinked around individuals' subjective life expectancy, i.e. such expectations about length of life are a reference point (van Nooten & Brouwer, 2004, van Nooten et al., 2009). However, the focus in this paper is to extend RP preferences to health, and hence, we will not extensively test or discuss the *assumptions* that generate the paradox. Furthermore, although such kinked preferences around subjective life expectancy may invalidate the assumptions necessary to generate RP, they increase the need to consider reference-dependent models for decisions about health. The limited evidence that we obtained to sustain Rabin's (2000) empirical assumptions is discussed in the Online Supplements of this dissertation.

⁷ The definitions used here rely on strict preference (as our experiment only involves direct choices), but as shown in the Online Supplements of this dissertation the following preferences also constitute RP: $I_h \sim \mathcal{G}_{0.5}\ell$ and $I_h \preceq \mathcal{G}_{0.5}\mathcal{L}$.

Stimuli: The exact scenarios for all 6 gamble-pairs can be found in Table 2.1, while instructions are reprinted in the Online Supplements of this dissertation. In accordance with Bleichrodt and colleagues (2019) we only asked subjects if they would accept this gamble, to which they could respond “Yes” or “No”.

Additional measures: We collected demographic information on age, gender, body-mass index (BMI), subjective health (0 – 100 scale from worst to best imaginable health) and happiness (1-10 scale from completely dissatisfied to completely satisfied with life as a whole).

Table 2.1. Scenarios for Rabin Paradox (RP) gamble-pairs for individual and societal outcomes

Gamble-pair	Scenario	Outcome
<i>Individual</i>		
RP1	Imagine that it is possible to take a gamble that affects your remaining lifetime (e.g. living until 87). The outcome is added or deducted from your lifetime.	Hours
RP2	Imagine that you are 75 and will live with slight mobility problems (not able to walk more than 3 kilometers). You can gamble to change your lifetime (longer or shorter).	Hours
RP3	Imagine you are 75 and will live until 85 with light back pain (e.g. treatable with mild painkillers). You can gamble to change your life time.	Hours
<i>Societal</i>		
RP4	Imagine a chronic disease, which leads to considerable losses in quality and length of life. Normally this disease affects about 300,000 people in the Netherlands (e.g. cancer). A risky drug is developed, which may either increase the amount of cases or decrease the amount of cases.	Cases averted /extra cases
RP5	Imagine an outbreak of a fatal disease occurred. The disease will lead to considerable lives lost. You are considering to take a gamble, in which either 11 lives are saved or 10 additional lives are lost.	Casualties saved / extra casualties
RP6	Imagine you have the chance to obtain extra healthy life years for society, be means of an easy to implement, costless, medical procedure. As a reminder: you do not know to whom these life years will be distributed. The procedure also has a chance of resulting in a reduction of healthy life years for society.	Life years

Note: Each gamble-pair had the following forms, with numbers referring to different health outcomes depending on the pair: a) Moderate Stake Gamble $\mathcal{G}_p \ell$: (+11 , 0.5, -10), b) Calibrated Gamble $\mathcal{G}_p \mathcal{L}$: (+10.000, 0.5, -100)

Results

As can be seen from Table 2.2, for all items a small majority of the sample rejected the gambles for moderate stakes, while a large majority generally accepted calibrated gambles. These proportions were all significantly larger than 50% ($\chi^2(1, N = 201) > 6.81, p's < .009$) for all items but RP1 ($\chi^2(1, N = 201) = 1.44, p = .23$). Next, for each RP gamble-pair we determined how many subjects showed RP preferences (see Table 2.2). Out of all 4 possible preference patterns within gamble-pairs, RP preferences occurred most frequently (43% - 64%). However, a substantial part of the sample showed preference combinations consistent with EU by rejecting or accepting both gambles (individual: 13% and 39%, societal: 15% and 23%). Of all choices consistent with RP preferences a larger share (358 out of 632, i.e. 56%) occurred for societal outcomes ($\chi^2(1, N = 632) = 11.16, p < .001$). Inversely, the proportion of our samples' choices satisfying EU was smaller (227 out of 541, i.e. 42%)⁸ for societal outcomes ($\chi^2(1, N = 541) = 13.99, p < .001$). We also qualified these results with mixed logistic regression (see the Online Supplements of this dissertation), which suggested that RP preferences were more frequent for societal outcomes after controlling for the demographics collected (as described in: 'Additional measures').

Next, we explored to what extent RP preferences were stable within-subjects, by calculating what proportion of our sample showed this combination of preferences across gamble-pairs. As can be seen from Table 2.3, overall RP preferences were observed frequently, with the percentages of those showing RP preferences for all three gamble-pairs being near equal for individual and societal outcomes. When considering the stability of these preferences between domains, it appeared that many individuals that had no RP preferences for individual outcomes did show RP preferences for societal outcomes. A series of analyses in the Online Supplements of this dissertation shows that these preferences were more consistent and stable than would be expected if they were generated by a population satisfying EU or being completely indifferent (i.e. noise). Furthermore, RP preferences were more consistent than would have been expected if all choices were made independently across all gamble-pairs.

⁸ The remaining 2% of all choices over gamble-pairs consisted of accepting the moderate stake gamble, but turning down the calibrated gamble. Such preferences occurred for a negligible part of the sample and are not captured by RP preferences or EU. We will not discuss these counter-intuitive preferences in more detail.

Table 2.2. RP gamble-pairs with number of acceptances (acc.) vs. rejections (rej.) for moderate stake gambles (MSG, in columns) and calibrated gambles (in rows), with row and column totals (tot.)

Individual setting		RP1-MSG $\mathcal{G}_{0.5\mathcal{L}}$			RP2-MSG $\mathcal{G}_{0.5\mathcal{L}}$			RP3-MSG $\mathcal{G}_{0.5\mathcal{L}}$		
Calibrated Gambles		Rej.	Acc.	(Tot.)	Rej.	Acc.	(Tot.)	Rej.	Acc.	(Tot.)
RP1-RP3 $\mathcal{G}_{0.5\mathcal{L}}$	Rej.	15	3	(18)	35	8	(43)	26	4	(30)
	Acc.	94 ⁺	89	(183)*	87 ⁺	71	(158)*	93 ⁺	78	(181)*
	(Tot.)	(109)	(92)		(123)*	(79)		(119)*	(82)	
Societal setting		RP4-MSG $\mathcal{G}_{0.5\mathcal{L}}$			RP5-MSG $\mathcal{G}_{0.5\mathcal{L}}$			RP6-MSG $\mathcal{G}_{0.5\mathcal{L}}$		
Calibrated Gambles		Rej.	Acc.	(Tot.)	Rej.	Acc.	(Tot.)	Rej.	Acc.	(Tot.)
RP4-RP6 $\mathcal{G}_{0.5\mathcal{L}}$	Rej.	25	8	(33)	14	3	(17)	49	7	(58)
	Acc.	119 ⁺	49	(168)*	127 ⁺	57	(184)	112 ⁺	33	(145)*
	(Tot.)	(144)*	(57)		(141)*	(60)		(161)*	(40)	

Note: ^a RP preferences are signified by ⁺, * indicates the total proportion is significantly larger than 50%, by a pairwise χ^2 -test with $p < 0.05$

Table 2.3. Frequency (N) and proportion (%) of RP preferences counts (C) within-subjects

N (%)		Societal				Total individual
		C = 0	C = 1	C = 2	C = 3	
Individual	C = 0	19 (9%)	21 (10%)	22 (11%)	22 (11%)	84 (42%)
	C = 1	2 (1%)	6 (3%)	9 (4%)	8 (4%)	25 (12%)
	C = 2	4 (2%)	5 (2%)	9 (4%)	9 (4%)	27 (13%)
	C = 3	8 (4%)	12 (6%)	18 (9%)	27 (13%)	65 (32%)
Total societal		33 (16%)	44 (22%)	58 (29%)	66 (33%)	

Discussion

The goal of this study was to supplement the empirical literature on RP, by extending this classic critique of EU to the health domain. We replicate RP for health; that is, we observe risk aversion for gambles over moderate health stakes, which implausibly (and incorrectly for a majority of our sample) suggests that calibrated large stake gambles should also be turned down according to EU. These findings are in accordance with the two other empirical studies testing RP preferences in the monetary domain (Bleichrodt et al., 2019, Cox et al., 2013). Several different hypothetical health outcomes and contexts were used, where RP preferences were more pronounced for societal outcomes. To our knowledge, our study is one of the first finding risk aversion for moderate individual health outcomes⁹, with another example being Breyer and Fuchs (1982) who consider gambles over days with a 2 hour headache. Risk aversion for larger individual health outcomes, for example in the range of 0.5 to 20 years of life is observed frequently (e.g. Attema et al., 2013, Attema et al., 2016, Galizzi et al., 2016c, Oliver, 2018, van der Pol and Ruggeri, 2008), albeit these studies used a different methodology (i.e. certainty equivalences). For societal outcomes, studies have, for example, found risk aversion for life years (Eraker and Sox, 1981) or lives (Kemel and Paraschiv, 2018).

However, a substantial part (30-52%) of our sample did not violate EU by accepting or rejecting both gambles, which is similar to that observed in the only direct test of RP in the economic literature (Cox et al., 2013). A direct comparison to Bleichrodt et al. (2019) is not possible, as they only tested risk aversion for small stakes, but the proportion of their sample that accepts small stake gambles is lower compared to our sample. A surprising and unique result of our study is that a small set of the sample rejects both gambles, and the strong concavity these preference imply could be considered absurd (Wakker, 2010). Whereas earlier work on RP focused on criticizing the assumptions generating RP (e.g. Andersen et al., 2011, Harrison et al., 2017), or explaining its paradoxical nature (e.g. via reference-

⁹ We do not refer to our stimuli as small stake gambles, as we object to labeling any health loss as small, especially when our gambles concern human lives.

dependence, Bleichrodt et al., 2019), our work suggests that for a non-negligible group of individuals no paradox may exist to begin with. Nonetheless, turning down the opportunity of gaining over a year of life or saving 10,000 lives when risking moderate losses in lifetime or human lives seems difficult to justify. Whereas loss aversion may explain RP preferences, as Rabin (2000) suggested and Bleichrodt and colleagues (2019) established, it is straightforward to demonstrate that to explain acceptance of both gambles loss aversion would need to be extreme. Hence, we offer two explanations for these preferences not related to risk aversion. First, especially relevant to individual outcomes, some individuals may not be willing to live any longer in the reduced health states (in scenario 2 and 3). Such preferences are observed frequently for health states more severe than those under consideration here (i.e. maximum endurable time, Sutherland et al., 1982). A second explanation is that one may prefer not to take any gamble at all, out of the well-known preference for inaction over action when risking adverse outcomes (i.e. omission bias, Spranca et al., 1991).

Some additional methodological limitations deserve mentioning. First, this study was not specifically designed to test the validity of the assumptions present in Rabin's (2000) calibration theorem. Given that some of these have been challenged in the economic domain (e.g. Andersen et al., 2011, Harrison et al., 2017), this provides opportunities for future work. For example, it could be determined if risk aversion indeed holds for many (if not all) levels of initial health, either by testing this for a single individual at many (hypothetical) ages, or by comparing risk aversion between individuals with different ages, that are otherwise similar. Second, this study used a relatively small, homogeneous, convenience sample, which may limit its external validity. Nonetheless, it is common to start a new experiment in convenience samples, and extend it afterwards to representative samples. Third, our study relied on hypothetical scenarios without real incentives. Although the importance of incentive-compatibility for behavioral experiments in health has often been stressed (e.g. Galizzi and Wiesen, 2018), our goal of offering calibrated gambles in terms of health made such a procedure impossible. Furthermore, some evidence exists in the economic domain suggesting that risk preferences are not qualitatively different between hypothetical and incentive-compatible gambles, although they may be more variable in the former (for reviews, see: Camerer and Hogarth, 1999, Hertwig and Ortmann, 2001). Finally, our definition of RP and method only allowed for strict preferences, whereas the small differences in rates of acceptance and rejection for individual gamble-pairs suggest that part of our sample may actually have been indifferent for moderate stakes. However, if this was the case we would have observed less within-subjects stability and importantly such indifferences would still yield RP, as indifference for moderate stake gambles still implies risk aversion, and thus strong concavity under EU (see the Online Supplements of this dissertation).

Conclusion

This study has shown that the paradox proposed by Rabin (2000) is also relevant to health outcomes. Given its large impact in economics, its implications for health deserve further study. It poses a challenge to earlier work in health economics which described risk aversion

by means of EU with concave utility over health outcomes, as this would lead to implausible conclusions for the large stakes often present in the health domain—this is the main message of Rabin (2000). This appears to hold especially for models describing societal decision-making. Fortunately, by modelling preferences (over health outcomes) within a reference-dependent framework such as prospect theory (Tversky and Kahneman, 1992), RP can be easily resolved. The increasing attention for such reference-dependent frameworks in health economics (Abellan-Perpiñan et al., 2009, Attema et al., 2013, Pinto-Prades and Abellan-Perpiñan, 2012, and also Chapter 7 of this dissertation) in work seeking more accurate descriptive theories is supported by our findings. Although these reference-dependent theories may be general enough to capture the strong risk aversion demonstrated by a small part of our sample, further investigation to understand how we decide about health under risk is clearly still needed.

3

A QALY LOSS IS A QALY
LOSS IS A QALY LOSS:
A note on independence
of loss aversion from
health states

Chapter based on:

Lipman, S.A., Brouwer, W.B.F., & Attema, A.E. (2019). A QALY loss is a QALY loss is a QALY loss: a note on independence of loss aversion from health states. The European Journal of Health Economics, 1-8.

Abstract: Evidence has accumulated documenting loss aversion for monetary and, recently, for health outcomes – meaning that generally losses carry more weight than equally-sized gains. In conventional Quality-Adjusted Life Year (QALY) models, which comprise utility for quality and length of life, loss aversion is not taken into account. When measuring elements of the QALY model, commonly the (implicit) assumption is that utility for length and quality of life are independent. First attempts to quantify loss aversion for QALYs typically measured loss aversion in the context of life duration, keeping quality of life constant (or vice versa). However, given that QALYs are multi-attribute utilities, it may be possible that the degree of loss aversion is dependent on, or inseparable from, quality of life and non-constant. We test this assumption using non-parametric methodology to quantify loss aversion, under different levels of quality of life. We measure utility of life duration for four health states within-subjects, and present the results of a robustness test of loss aversion within the QALY model. We find loss aversion coefficients to be stable at the aggregate level, albeit with considerable heterogeneity at the individual level. Implications for applied work on prospect theory within health economics are discussed.

Introduction

Like other decisions, medical decisions often involve trade-offs between gains and losses in different domains. In health economics, an important trade-off concerns that between length and quality of life (QoL), also in the context of health state valuations. Research in behavioral economics and psychology has established that in such trade-offs losses typically carry more weight than gains of the same size. This sensitivity to losses is referred to as *loss aversion* (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992). Recently, scholars demonstrated the importance of loss aversion within the health domain, both for life duration (Attema et al., 2013, Bleichrodt and Pinto, 2002, Oliver, 2003a, and Chapter 7 of this dissertation) and quality of life (QoL) (Attema et al., 2016, Bleichrodt and Pinto, 2002, Stalmeier and Bezembinder, 1999). In health economic analyses, utilities are often defined as a product of these two attributes, jointly comprising Quality-Adjusted Life-Years (QALYs) (Pliskin et al., 1980). Commonly, the utility function over these two outcomes is decomposed into separate utility functions over life duration and QoL. This separability of QALYs is, however, only possible under several assumptions, which have solely been tested under conditions in which no distinction is made between gains and losses (Bleichrodt et al., 2009).

Here, we use prospect theory (PT), which incorporates loss aversion and judges changes from the perspective of some relevant reference-point (RP). Bleichrodt and colleagues (2009) established that when considering multi-attribute outcomes, such as QALYs, gains and losses may be determined per attribute with separate attribute-specific RPs. This also makes it possible quantify loss aversion, to see how much more weight losses carry than gains. Earlier attempts at quantifying loss aversion under PT have typically focused on single attributes within the QALY framework, for example by obtaining loss aversion for life duration while maintaining QoL constant (Attema et al., 2013, see also Chapter 7 of this dissertation) or vice versa (Attema et al., 2016). Although these studies produced similar median estimates of loss aversion, with health losses receiving between 1.5 and 2 times more weight than gains, they did not allude to the issue of separability. In other words, these studies ignored the possibility that loss aversion for one attribute (e.g. length of life) depends on the level of the other attribute (which is typically held constant) and, hence, assumes loss aversion for health outcomes to be constant, independent of their QALY profile.

However, it could be the case that some QALY losses carry more weight relative to commensurate QALY gains than others, for example if loss aversion is more pronounced for more severe health states. In this article we test this assumption using a non-parametric method (Abdellaoui et al., 2016) to quantify loss aversion over life duration, under varying levels of QoL. This non-parametric method was developed recently and allows estimation of utility curvature and loss aversion without imposing parametric assumptions on either. Earlier work has argued that the choice of parametric family or functional form restricts interpretation of subjects' choice patterns, and may lead to considerable bias especially for extreme cases (Abdellaoui, 2000, Abdellaoui et al., 2016). This method was first adapted to and used in the health domain for the study reported in Chapter 7 of this dissertation.

Theoretical framework

Consider a decision maker facing choices with regard to his health under uncertain conditions, operationalized by presenting decision makers with prospects representing different life durations and QoL. We assume completeness and monotonicity for both attributes. We consider lotteries involving chronic health profiles, described as (β, T) , where β represents QoL and T duration in years. According to the generalized QALY model (Miyamoto and Eraker, 1989), a decision maker's preferences for health profiles can be represented by:

$$V(\beta, T) = U(\beta) * L(T), \quad (3.1)$$

with $V(\beta, T)$ being a product of $U(\beta)$, the utility of β , and $L(T)$ denoting the utility of T life years.

Here, we assume PT with a sign-dependent utility function for life duration, so that gains are evaluated differently than losses, relative to an attribute specific reference point. We assume that through instruction, it is possible to set this attribute-specific RP to a specific health condition β_c and life duration T_0 . In order to elicit a continuous utility function for life duration, we elicit a standard sequence for life duration that runs through $L(T_0) = 0$. Meanwhile, we keep QoL constant at β_c throughout the task. We repeat this process under different levels of β_c .¹⁰

We elicit the utility function for life duration, relative to this reference point, both for gains and losses for the different health states. Hence, we obtain $L^i(T)$ for each β_c , 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 $L^i(T_0) = 0$. We incorporate loss aversion by taking $L^-(T) = \lambda L(T)$ for $T < T_0$, where λ denotes a loss aversion index, with $\lambda > 1$ [$= 1, < 1$] indicating loss aversion [loss neutrality, gain seeking]. Hence, by obtaining the utility around the reference point, the degree of loss aversion can be derived.

Methods

A total of 111 students (average age: 20.23, SD = 1.52) of Rotterdam School of Management (61 female) participated in this study for a course credit reward. Experimental sessions lasted for 25 minutes and were run with up to 4 subjects per session. One experimenter was present in the room to answer questions. The experiment was computerized with Matlab.

To test the robustness of loss aversion, we used the non-parametric method (Abdellaoui et al., 2016) under four levels of QoL. In other words, each subject completed the non-parametric method four times, with a different β_c throughout each of these four phases. This process allows us to obtain estimates of utility curvature and loss aversion for each of the four levels of QoL, and compare them within-subjects.

QoL was defined by means of EQ-5D-5L health state descriptions (Herdman et al., 2011), which utilize five domains: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The 5L version of the EQ-5D distinguishes five levels of severity on each

¹⁰ For a classification of PT under these assumptions see Chapter 7, and for a more elaborate description of additive utility under PT, see Bleichrodt et al. (2009).

domain, ranging from ‘no problems’ to ‘extreme problems/unable to’. Health states are typically denoted by 5 digit codes like 22113, with each number representing severity of the relevant domain level of QoL. In this study, we used four relatively mild to moderate health states as RP (β_c) in the non-parametric method: 11111, 21211, 31221, and 32341 (see the Online Supplements of this dissertation for exact description). This was done to have variation in health states but avoid states worse than dead, for which no separate procedure was included.

The non-parametric method used here consisted of three stages which are described in detail in the Online Supplements of this dissertation.¹¹ The first stage connects the utility for gains and losses. The second and third stages employ the trade-off method developed by (Wakker and Deneffe, 1996) to measure a standard sequence of outcomes in life years for gains ($x_1^+, x_2^+, \dots, x_5^+$), and for losses ($x_1^-, x_2^-, \dots, x_5^-$). This enables measuring loss aversion, without imposing parametric assumptions on utility curvature.¹² Additionally, standard sequences allow the testing of utility independence (Bleichrodt et al., 2009). The three stages had slightly different instructions, providing context for the required trade-offs. The instructions were similar to those used in Chapter 7 of this dissertation. During all stages of the experiment, it was made clear to subjects that they should imagine living until 70 years in β_c , after which they would contract a disease resulting in immediate death without any pain. Subjects completed a series of binary choices between two drugs which could change their situation (leading to gains and losses compared to living until 70). Employing a bisection choice method, we obtained indifferences, set equal to the midpoint after the 5th binary choice. Some stimuli and constants relevant to the non-parametric method had to be set beforehand; these are listed in the Online Supplements of this dissertation.

Results

Seven subjects were excluded from further analyses for the following reasons: mechanical failure ($n = 2$), refusing to incur life year losses ($n = 3$), and observed misbehavior (e.g. rushing through the task, $n = 2$). The results are reported for the reduced sample ($n = 104$)¹³. Throughout, we will first report aggregate analyses, where median parameters are compared for the whole sample, and refer to these as results at ‘the aggregate level’. Second, we will investigate individual results more closely, by classifying each individual according to classification rules reported in Box 3.1 and we explore within-subjects parameter instability. We refer to these analyses as ‘individual-level analyses’.

Table 3.1 demonstrates the results at the aggregate level, by comparing point-estimates for utility curvature and loss aversion for each health state. We compared differences between health states using omnibus tests (i.e. comparing all four health states simultaneously), more specifically Friedman’s tests, which are robust against the violations of normality typically observed for parameters under the definitions reported in Box 3.1. Next, we compared all health states in pairs with Wilcoxon signed ranks tests. For the omnibus tests, no significant

¹¹ For an elaborate, formal description of this method, see Abdellaoui and colleagues (2016).

¹² For more information on how utility curvature and loss aversion were determined, see Box 3.1.

¹³ Conventional post-hoc power analyses suggested this sample was sufficiently powerful to enable detecting differences with at least small effect sizes (Cohen’s $d < 0.3$), assuming $\alpha=0.05$ and statistical power at the recommended 80% level (Cohen, 1988).

differences were observed between health states, both for utility curvature and loss aversion (all p 's $> .06$). When comparing parameter estimates in pairs of health states, some significant differences were observed. For loss aversion under both definitions, parameter estimates for β_2 were significantly lower than for β_3 (p 's < 0.03). All other pairwise comparisons for loss aversion yielded no significant differences (all p 's > 0.07). Using pairwise comparisons for utility curvature we observe no significant differences for both parametric and non-parametric estimations (all p 's > 0.05).

In general, we observe close to linear utility for all health states, both for gains and losses¹⁴. Furthermore, we observe considerable loss aversion at the aggregate level, with λ significantly greater than 1 for all β_c (Wilcoxon tests: $p < .001$ for all β 's).

Table 3.1. Median (IQR in brackets) parameter point-estimates for loss aversion under two definitions and utility curvature as defined by area-under-curve (AUC) and power utility

Health state	β_0 : 11111	β_1 :21211	β_2 : 31221	β_3 : 32341
Utility curvature				
AUC – Gains	0.51 (0.42 – 0.63)	0.49 (0.38 – 0.59)	0.53 (0.44 – 0.64)	0.52 (0.41 – 0.70)
AUC – Losses	0.51 (0.46 – 0.57)	0.50 (0.45 – 0.57)	0.50 (0.42 – 0.58)	0.49 (0.40 – 0.60)
Power – Gains	0.96 (0.58 – 1.37)	1.07 (0.69 – 1.71)	0.91 (0.57 – 1.28)	0.78 (0.45 – 1.41)
Power – Losses	0.93 (0.74 – 1.16)	0.94 (0.73 – 1.20)	0.97 (0.73 – 1.41)	1.02 (0.66 – 1.40)
Loss aversion				
Köbberling Wakker	1.97 (1.33 – 4.43)	1.93 (1.45 – 3.67)	1.88 (1.39 – 3.30)	2.13 (1.15 – 8.38)
Kahneman Tversky	2.13 (1.24 – 4.39)	1.94 (1.26 – 4.62)	2.10 (1.25 – 3.23)	2.51 (1.18 – 6.24)

Table 3.2 demonstrates how subjects classify under different estimations of utility curvature and loss aversion (see Box 3.1). For all individual classifications, we observed that the conventionally assumed loss neutrality and linear utility curvature are not present in our data. Although at the aggregate level linear utility was found, when classifying individually, considerable heterogeneity in utility curvature was observed, with proportions of concave/convexity varying between definitions and health states. This finding could be explained by the near equal division of concavity/convexity in our sample, resulting in roughly linear utility at the aggregate level. For loss aversion, however, such an equal division was not visible, with the majority of subjects classifying as loss averse across definitions and health states.

¹⁴ Wilcoxon tests comparing non-parametric curvature estimates with AUC=0.5, and parametric estimates with $\alpha = 1$, produced no significant results for all β (all p 's > 0.08), with one exception: β_1 power utility for gains, $p=0.04$.

Table 3.2. Individual classifications for utility curvature (n = concave, linear, convex), and loss aversion (n = loss averse, loss neutral, gain seeking)

Health state	β_0 : 11111	β_1 :21211	β_2 : 31221	β_3 : 32341
Utility curvature				
AUC – Gains	55, 0, 49	47, 0, 54	61, 0, 43	61, 0, 43
AUC– Losses	44, 0, 60	42, 0, 62	49, 0, 55	56, 0, 48
Power – Gains	54, 0, 50	47, 0, 57	62, 0, 42	65, 0, 39
Power – Losses	41, 0, 63	41, 0, 63	51, 0, 53	53, 0, 51
Loss aversion				
Köbberling/Wakker	90, 0, 14	92, 0, 12	95, 0, 9	89, 0, 15
Kahneman/Tversky	86, 0, 15	85, 0, 17	89, 0, 13	82, 0, 18

Our design allowed exploring point-estimate stability for utility curvature and loss aversion between different levels of β_c . To this end, we calculated the difference between the smallest and largest estimate within-subjects (e.g. the lowest and highest λ). Furthermore, to allude to within-subjects heterogeneity in classification, we calculated the proportion of subjects for whom classifications were dependent on health states (e.g. loss averse for β_0 to β_2 , gain seeking for β_3). Both exploratory measures of within-subjects parameter and classification variance demonstrated considerable heterogeneity between health states (see Table 3.3). Finally, we investigated whether systematic patterns in utility curvature or loss aversion could be observed in our sample. To this end, we determined the extent to which subjects showed monotonically increasing (or decreasing) parameters (see Table 3.3). For loss aversion, this classification indicated that subjects became more (less) loss averse for increasing health state severity for β_c . These analyses indicate that these patterns did occur, but only for a small part of our sample, again suggesting non-systematic heterogeneity of parameter estimates.

Table 3.3. Exploration of within-subjects heterogeneity for different health states

	Parameter point-estimate difference (max – min)	Health-state dependent classifications (%)	Monotonically increasing/ decreasing (%)
Utility curvature			
Non-par – Gains	0.26	75%	6% / 7%
Non-par – Losses	0.18	76%	4% / 7%
Power – Gains	1.17	73%	7% / 7%
Power – Losses	0.75	79%	5% / 5%
Loss aversion			
Köbberling Wakker	5.10	37%	8% / 7%
Kahneman Tversky	4.17	49%	6% / 7%

Discussion

In this paper we compared estimates for utility curvature and loss aversion for QALY outcomes under four levels of QoL, to test the robustness of these estimates. An extensive literature exists testing the validity of QALY models, which has documented mixed evidence with regard to the separability of life duration and QoL (e.g. Abellan-Perpignan et al., 2006, Attema and Brouwer, 2012b, Bleichrodt and Pinto, 2005, Miyamoto and Eraker, 1988). Additionally, many authors have investigated utility independence with regard to health state valuation (e.g. the relation between utilities and time horizon in standard gambles), finding many descriptive violations of this independence (for a review, see: Attema and Brouwer, 2012b). Ours was the first experimental test of this separability for QALY gains and losses separately, and we also tested the robustness of loss aversion. Our results, at the aggregate level, provided evidence that estimations of loss aversion and utility curvature are independent of QoL. However, loss aversion and utility curvature estimates were heterogeneous at the individual level, i.e. varied considerably between health states for the same individual.

Our findings are in many regards similar to earlier work that measured PT for QALY outcomes. We observed considerable loss aversion (defined over length of life), as was found in similar magnitude in earlier work applying similar methodology (Attema et al., 2018a, and Chapter 7 of this dissertation), or with different elicitation methods (Attema et al., 2013, Attema et al., 2016). In contrast to what was observed in earlier applications of the non-parametric method for health outcomes (Attema et al., 2018a, and in Chapter 7 of this dissertation), we found linear utility for both gains and losses at the aggregate level. Applying a parametric approach to our non-parametric measurements did not affect these conclusions. However, when estimating individual classifications, we found none for whom our data

supported this linearity, as we observed a near equal spread in concave/convex utility (i.e. averaging out to linear).

We document considerable heterogeneity in parameter estimates between-subjects, and in also observed such heterogeneity within-subjects for different health states. Our exploratory analyses did not uncover systematic or monotonic patterns in this within-subjects heterogeneity. An explanation related to our chosen chained utility elicitation method could be that these individual differences occurred as a result of preference imprecision (Bhatia and Loomes, 2017). Such ‘noisy preferences’ could result in error propagation, i.e. cascading of errors or imprecision in early stages of our chained method into later stages, producing differences in parameters between health states when errors occur randomly. Although earlier work using similar methodology (Bleichrodt and Pinto, 2000, and Chapter 7 of this dissertation) observed no effects of error propagation, we cannot rule out it affected current results. Another factor contributing to possible error propagation in our study could be that we opted to obtain indifferences via bi-section only (to reduce complexity), whereas earlier work (Abdellaoui et al., 2016) using this method applied a slider to obtain indifferences, allowing subjects to correct errors adaptively. Future work could explore this further, for example by adding a slider to obtain indifference points, using non-chained methodology, or running an error propagation simulation.

Some additional limitations of this study deserve noting. First, since this study involved a first test of independence of loss aversion in health, we used a convenience sample consisting of students. Of course, future extensions preferably should include representative samples to generalize our findings. Although power analyses suggested that our sample was adequately powered to detect small effects, using a larger sample could perhaps result in the detection of smaller effects, also given the large heterogeneity for parameter estimates reported here. Second, we assumed it is possible to set the RP through instruction, while it may be the case that respondents took another RP in mind. Still, given the high loss aversion coefficients we found, it seems plausible that our respondents indeed held the induced RP in mind. Finally, our study used four mild to moderate health states, including perfect health, while the EQ-5D descriptive system enables many more possible health states, with more severe health problems than our selection. Given the aim of our study this is a clear limitation, as perhaps these states were insufficiently spaced in terms of utility for us to observe systematic patterns in loss aversion on utility curvature parameters. However, our empirical approach required us to make a fundamental assumption: monotonicity. The non-parametric method breaks down if monotonicity is not satisfied, i.e. if subjects prefer to lose years of life instead of gaining them. For more severe health states, monotonicity need not always hold (Sutherland et al., 1982). Obviously, many other mild health states were available for our purposes, but to reduce cognitive strain for our subjects we decided on including just four. For reference, these four health profiles receive utility weights ranging from 1.00 to 0.46 in the Dutch tariff (Versteegh et al., 2016), which we considered to be sufficient for our purposes. Future work could replicate our findings with a different or larger selection of health states.

Our findings may have implications for policy makers and researchers aiming to apply PT measurements to health-related decision-making. Our results imply that median parameters in applications of PT may have merit, as these estimates appear to be robust across different scenarios (in terms of QoL). For example, our work warrants the conclusion that, at the

aggregate level, life year losses are weighed twice as much as similarly-sized gains, regardless of QoL level. However, as our exploratory analyses of within-subject heterogeneity demonstrated, individuals' loss aversion and utility curvature may depend on the health state used during elicitation. This heterogeneity at the individual level may be problematic for approaches using averages, like median-optimized parameters (Pinto-Prades and Abellan-Perpiñan, 2012). When aiming to address such loss aversion at the individual level, our data would suggest that assuming such median loss aversion parameters may misrepresent individuals' actual preferences and trade-offs. When one aims to apply PT to allude to biases in individual cases (e.g. in health state valuation), an individual approach may be more suitable, given both the considerable between-subjects and between-health states heterogeneity reported in this study. Such corrections with individually estimated parameters could be too time consuming and labor-intensive when applied separately for each economic evaluation. However, in many countries, such as the UK, QALYs are not derived individually, but from indirect' preference-based classification systems, such as EQ-5D or SF-6D via tariff lists (Drummond et al., 2015). Recent developments in de-biasing QALY measurement (see Chapter 7 and 8 of this dissertation) have created new opportunities and arguments in favor of individual correction for PT within these frameworks, where proponents argue for large scale valuation studies to include (individual) corrections (see Chapter 7 and 8 of this dissertation). Given that these valuation studies aim to obtain accurate QALY weights, it seems important to consider which health state is used to quantify PT parameters, if such individual correction is to be applied.

In conclusion, although we observed large heterogeneity of loss aversion and utility of life duration depending on QoL, we failed to observe systematic patterns in this dependence, and observed no differences on average. Future work should aim to address whether this heterogeneity is method-dependent or due to systematic differences between individuals or health states. For now, it appears that on average loss aversion is equal across health states, i.e. a QALY loss is a QALY loss is a QALY loss, and it receives approximately twice as much weight as equally sized QALY gains.

Box 3.1. Analyses of utility curvature and loss aversion

We non-parametrically calculated the area under the curve for $L^i(T)$, which was normalized to $[0,1]$, for gains and $[0, -1]$ for losses. If utility is linear, the area under this normalized curve equals one half for both gains and losses. Utility for gains in life duration is convex (concave) if the area under the curve is smaller (larger) than one half, while for losses the opposite direction holds (convex $> 1/2$, concave $< 1/2$). Second, we fitted a parametric utility curve to our data by employing the power family, with the utility of life duration defined as x^α 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.

Kahneman and Tversky (1979) defined loss aversion (λ) as $-U(-x) > U(x)$ for all $x > 0$. To measure loss aversion coefficients according to this definition, we computed $-U(-x_j^+)/U(x_j^+)$ and $-U(-x_j^-)/U(x_j^-)$ for $j = 1, \dots, 5$. As a result of the trade-off procedure, $U(-x_j^+)$ and $U(-x_j^-)$ could usually not be observed directly and thus were determined through linear interpolation. Subjects were classified as loss averse if $-U(-x)/U(x) > 1$ for more than half of the observations, as loss neutral if $-U(-x)/U(x) = 1$ for more than half of the observations, and as gain seeking if $-U(-x)/U(x) < 1$ for more than half of the observations.

Köbberling and Wakker (2005) provided an easier method to determine loss aversion. They defined loss aversion (λ) as the kink of utility at the reference point. That is, they defined loss aversion as $U'_1(0)/U'_2(0)$, with $U'_1(0)$ representing the left derivative and $U'_2(0)$ the right derivative of U at the reference point. To operationalize this definition, we computed each subject's coefficient of loss aversion as the ratio of $U(x_1^-)/x_1^-$ over $U(x_1^+)/x_1^+$, because x_1^- and x_1^+ are the loss and gain elicited closest to the reference point. 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 $x_1^+/-x_1^- < 1$.

4

One size fits all? Designing financial incentives tailored to individual preferences

Chapter based on:

S.A. Lipman (2020). One size fits all? Designing financial incentives tailored to individual preferences. Behavioural Public Policy.

Abstract: Financial incentives are often designed to benefit from behavioral insights. Individuals' preferences for such behaviorally inspired incentives are rarely studied, nor is the role the behavioral insights that motivated them play. This study aimed to let individuals design their own incentives (i.e. tailored incentives) and explore which individual characteristics are associated with these preferences for tailored incentives. A sample of students ($n = 182$) tailored hypothetical incentives for visiting the gym. Incentives could be tailored by: a) committing personal funds, b) picking weekly pay-outs (increasing or decreasing), and c) introducing pay-out risk whilst increasing value. Afterwards, (inter alia) loss aversion, probability weighting, time discounting, present bias, cognitive reflection, and trait self-control were measured. A large majority indicated to be willing to deposit their own money, and only very few individuals select risky incentives. These heterogeneous preferences for financial incentives are poorly predicted by the individual characteristics measured (i.e. economic preferences and psychological traits). These results suggest that preferences for tailored incentives could be studied as input for the design of financial incentives. However, it is unclear if tailoring incentives improves cost-effectiveness, as the lack of association between tailored incentives and the behavioral insights that motivate them has multiple conflicting interpretations.

Introduction

Financial incentives appear to be a promising public policy tool to promote behavior change for the most prominent causes of chronic, non-communicable disease (WHO, 2009), such as tobacco use, poor diet and physical inactivity (for systematic reviews, see: Giles et al., 2014, Mantzari et al., 2015, Mitchell et al., 2013, Strohacker et al., 2013). Many different financial incentive schemes are used, which differ for example in terms of the size, timing or certainty of payment (Adams et al., 2014). Often, insights from behavioral economics are used to motivate or design the financial incentives used. For example, financial incentives have been used that capitalize from behavioral insights such as loss aversion (e.g. deposit/commitment contracts: Bhattacharya et al., 2015, Bryan et al., 2010, Giné et al., 2010, Volpp et al., 2008) or probability weighting (e.g. lottery incentives: Haisley et al., 2012, Kimmel et al., 2012, van der Swaluw et al., 2018, Volpp et al., 2008). The effectiveness of such financial incentives, which Galizzi (2014) refers to as behaviorally inspired incentives, is hypothesized to be amplified by deviations from traditional rationality.

However, no conclusive evidence exists to support policymakers in the choice between different (behaviorally inspired) financial incentive schemes. Several randomized control trials (e.g. Haisley et al., 2012, Patel et al., 2016) have systematically compared different incentive schemes directly against each other (e.g. lottery vs. commitment incentives), or against fixed incentives (Halpern et al., 2011). However, given the costly nature studying financial incentives, such studies far from exhaust all possible comparisons between behaviorally inspired incentives. The use of incentive schemes that mix behavioral components of different designs is even rarer (e.g. van der Swaluw et al., 2018). As a result, it is unclear who responds to financial incentives and why (Paloyo et al., 2015), which may explain why a one-size-fits-all approach is often applied: offering all respondents the same type of financial incentives (often motivated by a single behavioral insight, if any at all). The main motivation of this paper is to move beyond such one-size-fits-all approaches, and instead provide incentives tailored to individuals' preferences. This extends earlier work on incentives in two domains.

First, existing work mostly compares behaviorally inspired incentives by means of random assignment (e.g. Kullgren et al., 2016, Volpp et al., 2008), rather than exploring which types of financial rewards individuals prefer themselves. However, scarce evidence suggests that preferences for behaviorally inspired incentives are heterogeneous. For example, Halpern and colleagues (2015) find that only 14% voluntarily accept deposit contracts, while Vashishta and colleagues (2015) find a small majority prefers lotteries over fixed incentives (in a non-health context). Allowing individuals full autonomy in selecting those incentives they prefer (i.e. tailoring incentives) could increase individuals' motivation to engage in healthy behavior (see the work on self-determination theory by: Deci and Ryan, 2008). Hence, in this study, a newly-developed tool is implemented which individuals to tailor their own incentives, i.e. each individual could select a unique combination of different incentive design elements. This tool was tested in a lab experiment, in which individuals were asked to self-select (hypothetical) tailored incentives to promote exercise.

Second, even though large heterogeneity exists in the economic preferences motivate behaviorally inspired incentive designs, the importance of these individual differences has rarely been explored in the context of financial incentives. For example, a plethora of work in

experimental economics has shown large differences in for example probability weighting and loss aversion, both for money (e.g. Abdellaoui et al., 2016, Abdellaoui et al., 2007, Bleichrodt et al., 2016, Bruhin et al., 2010, Kahneman and Tversky, 1979, Tversky and Kahneman, 1992) and for health (e.g. Attema and Lipman, 2018, Kemel and Paraschiv, 2018, and Chapters 3 and 7 of this dissertation). This large heterogeneity in economic preferences raises several issues. For example, it is unknown if those who show preferences consistent with a behavioral insight (e.g. are loss averse) are to a larger extent affected by financial incentives designed with this behavioral insight in mind (e.g. deposit contracts) than those who are not. To date, only some evidence exists for lottery incentives for secondary prevention (Björkman Nyqvist et al., 2018) and financial incentives for exercise procrastination (Woerner, 2018). Furthermore, it is unknown if these economic preferences may explain heterogeneity in uptake of behaviorally inspired incentives, i.e. if those who are loss averse would be more (or less) likely to sign up for deposit contracts. This study addresses the latter issue, by for each respondent measuring a set of economic preferences, which are often used to motivate particular incentive design choices. To further explore who responds to financial incentives and why, the association between these economic preferences and tailored incentives is investigated.

Experiment

Sample and setting

The sample consisted of 182¹⁵ Business Administration students (63 females, average age = 19.17, SD = 1.47) who were rewarded course credits for their participation. Sessions lasted 30 minutes and were run in adjacent cubicles with an instructor present to answer any questions.

Tool for tailored incentives

Students were presented with a (hypothetical) scenario, in which their employer was facilitating their achievement of a weight-loss goal by offering a financial incentive for visiting the gym at least twice every week for a 10-week period. The reward had a fixed expected value of 100\$ over this 10-week period. The tool for tailored incentives¹⁶ allowed individuals to interactively design their own incentive scheme, whilst keeping the expected value of the incentives constant. The following instruction was used: *‘Your employer is quite flexible, and besides the expected pay-out has no preference in how your financial reward is structured. Obviously, you yourself know best what kind of pay-out structure would motivate you to go to the gym and reach your goal of losing weight. Therefore, we ask you to indicate how you would like your pay-out(s) to be structured’*. Students could tailor incentives along four dimensions, by: a) deciding to commit personal funds (*Pre-commitment*)¹⁷, b) picking weekly pay-outs (*Timing*), c) which could be increasing or decreasing (*Sequence*), and d)

¹⁵ To my knowledge, this is the first study of tailored incentives, i.e. no studies were available as a basis for a priori sample size calculation. Post-hoc power analysis suggests that this study was powered to find small to medium effects (see the Online Supplements of this dissertation for analysis and interpretation).

¹⁶ This tool was developed in Shiny, an R package allowing the development of web-apps. A demo version of the task is available at: <https://referencepoints.shinyapps.io/Minecentive/>. Code is available on request, and the task can be adapted by any researchers, organizations or policy makers interested in tailoring incentives.

¹⁷ Such incentives in which individuals commit their personal funds are often referred to as deposit contracts. In the tool designed for this study, this option was called Pre-committing.

introducing pay-out risk that increases value (*Risk*). Table 4.1 shows an overview of the framing and parametrization used for each dimension.

Economic preference elicitation

Table 4.2 provides an overview of the economic preferences elicited in this study, and the implications of these risk and time preferences (full detail on measurement and definitions can be found in Online Supplements). Risk preferences were elicited by measuring loss aversion, utility curvature (for gains and losses) and probability weighting (for gains and losses) using non-parametric methodology (adapted from Abdellaoui, 2000, Abdellaoui et al., 2016). This methodology has been recently introduced, and inter alia successfully applied to measure risk and time preferences for decisions about money and health (Abdellaoui et al., 2016, and Chapters 3 and 7 of this dissertation). The use of such non-parametric methodology may be preferred as it does not rely on certain parametric assumptions, which may not reflect preferences (Abdellaoui, 2000, Abdellaoui et al., 2007) or have troublesome mathematical properties around extremes (Wakker, 2008). Next, time preferences were measured assuming a quasi-hyperbolic discounting model (Laibson, 1997), where present bias (for gains and losses) and a weekly discount rate (for gains and losses) were elicited.

Exploratory questionnaires

Besides eliciting these economic preferences, subjects filled in a series of questionnaires, aimed at exploring the association between various psychological measures and tailored incentives. Several questions and questionnaires were used to measure self-reported health behaviors (alcoholic drinks/cigarettes consumed per week, exercise behavior and BMI), self-control (Tangney et al., 2018), cognitive reflection (Toplak et al., 2011), and personality (Francis et al., 1992) (reprinted in the Online Supplements of this dissertation). These questionnaires were only filled in by subjects in time remaining after they finished the main experiment, and hence not completed by all subjects (see Table 4.3 for number of complete observations per measure).

Results

All analyses are available on request and are reported without correcting for multiple testing.

Descriptive statistics

Table 4.2 (economic preferences) and Table 4.3 (demographics and psychological measures) show descriptive statistics for the sample. These results indicate that students generally are: non-smokers, moderate drinkers, engaging in regular exercise, loss averse, diminishingly sensitive to gains and losses, sensitive to extreme probabilities (i.e. inverse S-shaped probability weighting), present biased, and not or slightly discounting monetary amounts on a weekly basis. However, the standard deviations reported in Table 4.2 and 4.3 reflect the considerable between-subject heterogeneity that motivated this study.

Table 4.1. Overview of dimensions that could be edited to design tailored incentive schemes, with framing, options and parametrization

Dimension	Framing	Options
Pre-commitment	<p><i>'You can decide to pre-commit, by paying 100€ and your employer will add 100€. If you attain your weekly goals you will get this total amount of 200€, but you will lose (a part of) your committed 100€ if you don't attain it'</i></p>	<p><u>Do you want to pre-commit?</u> Yes, I will pay for entry No</p>
Timing	<p><i>'For each week that you attain your goal you will be rewarded. For example, if you attain your goals 8 out of 10 weeks, you will receive 80% of the reward. You can choose to receive all of your pay-out at the end of the 10 week period, or to receive parts of this sum in weekly parts for each week you attain your goal. Obviously, not attaining your goals will mean you do not receive any pay-out that week.'</i></p>	<p><u>How often should your pay-outs be?</u> One pay-out (at week 10) Weekly pay-outs</p>
Sequence	<p><i>'If you decide on weekly pay-outs, pay-out amounts can be fixed for each week, starting low or increasing or the other way around, the slider below lets you select different structures.'</i></p>	<p><u>What should your pay-off structure be?</u> (unbeknownst to respondents, each weekly reward was determined by multiplying the total available reward by a factor d_s, with: $d_s = (10 + b + r * t)/100, \quad t = 1, 2, \dots, 10$)</p>
Risk	<p><i>'Instead of receiving a sure amount, you may also receive your pay-out in the form of a lottery. Picking a lottery will increase your possible reward, but also increase the risk of not receiving any reward. The slider below lets you select different lottery structures.'</i></p>	<p>Slider options corresponded with the following parametrizations: Option 1: Strongly increasing, $b = -10, r = 2$ Option 2: Increasing, $b = -5, r = 1$ Option 3: Constant, $b, r = 0$ Option 4: Decreasing $b = 5, r = -1$ Option 5: Strongly decreasing $b = 10, r = -2$ <u>Chance of pay-out could be adjusted from:</u> $p = 1 - 100\%$. This led to a lottery which increased the (weekly) pay-out by a factor of $100/p$.</p>

Table 4.2. Elicited economic preferences (including median and interquartile range), with the implication of modal (i.e. most frequently occurring) preferences and related dimensions of tailored incentives

<u>Parameter</u>	<u>Median</u> <u>(Q1-Q3)</u>	<u>N (%)</u>	<u>N (%)</u>	<u>N (%)</u>	<u>Implication of modal preferences</u>	<u>Ref.</u>
Loss aversion (λ)*	1.61 (1.06 – 2.97)	$\lambda < 1$ 22 (12%)	$\lambda = 1$ 11 (6%)	$\lambda > 1$ 149 (81%)	Monetary losses carry more weight than equally sized gains	(Köberling and Wakker, 2005)
Utility curvature (α)	0.86 (0.58 – 1.11)	$\alpha < 1$ 114 (62%)	$\alpha = 1$ 7 (4%)	$\alpha > 1$ 61 (33%)	Each extra dollar gained carries less weight.	(Abdellaoui et al., 2016)
losses*	0.91 (0.70 – 1.18)	104 (57%)	7 (4%)	71 (39%)	Each extra dollar lost carries less weight.	(Abdellaoui et al., 2016)
Probability weighting (γ)	0.86 (0.76 – 1.39)	$\gamma < 1$ 123 (67%)	$\gamma = 1$ 4 (2%)	$\gamma > 1$ 55 (30%)	Small (large) chances of gains are overweighted (underweighted)	(Kahneman and Tversky, 1979)
losses*	1.00 (0.78 – 2.63)	88 (48%)	4 (2%)	90 (49%)	Small (large) chances of gains are overweighted (underweighted)	(Kahneman and Tversky, 1979)
Present Bias (β)	0.99 (0.91 – 1.00)	$\beta < 1$ 135 (74%)	$\beta = 1$ 5 (3%)	$\beta > 1$ 41 (22%)	Gains incurred now always carry more weight than those in the future.	(Laibson, 1997)
losses*	0.99 (0.93 – 1.01)	101 (55%)	17 (9%)	64 (35%)	Losses incurred now always carry more weight than those in the future.	(Laibson, 1997)
Discounting (δ)	0.01 (0.00 – 0.04)	$\delta < 0$ 21 (11%)	$\delta = 0$ 6 (3%)	$\delta > 0$ 155 (85%)	The positive value assigned to a dollar gained diminishes over time.	(Laibson, 1997)
losses*	0.00 (0.00 – 0.01)	66 (36%)	30 (16%)	86 (47%)	The negative value assigned to a dollar lost increases over time.	(Laibson, 1997)

Note: * indicates that that this distribution was different than that expected by chance, tested with Chi-squared tests and a significance level of 0.05. For definitions and implications of $\lambda, \alpha, \gamma, \beta$, and δ , see the Online Supplements of this dissertation).

Tailored incentives

The results of the tool for tailored incentives can be found in Table 4.4. A significant majority decided to pre-commit personal funds to increase rewards (Chi squared test, $\chi^2(1, N = 182) = 22.51, p < 0.001$), and a near-even split existed in the sample for preferences for one or weekly pay-outs. Those preferring weekly pay-outs, generally preferred slightly increasing or constant pay-outs. Lottery incentives were infrequently selected, with a negligible group (3 out of 163) selecting the lowest possible chance of winning and a large and significant majority preferring certain pay-outs rather than any of the other possible probabilities of pay-out (Chi squared test, $\chi^2(34, N = 182) = 1397, p < 0.001$). The three most prominent tailored incentive schemes were: pre-committing with one certain pay-out (12% of the sample), pre-committing with weekly, constant pay-outs (8% of the sample), and pre-committing with weekly, slightly increasing pay-outs (8% of the sample).

Association between selected tailored incentives and economic preferences

Next, a series of analyses was performed to explore the association between the tailored incentives students selected and their economic preferences.

First, these associations were explored by means of *t* tests (for *Pre-commitment* and *Timing* dimensions) and Spearman rank-correlation analyses (*Structure* and *Risk* dimensions) between individuals' choices on each dimension and the various measures obtained, which showed no consistent associations. For example, no significant differences were observed between individuals who chose deposit contracts or not for: loss aversion, probability weighting, utility curvature, present bias, time discounting, health behaviors, cognitive reflection, personality, and trait self-control (*t* tests, all p 's > 0.08). A similar lack of evidence for *Timing* (*t* tests, all p 's > 0.07), *Structure* (all Spearman ρ 's $< 0.15, p$'s > 0.08), and *Risk* (all Spearman ρ 's $< 0.11, p$'s > 0.15) can be observed. The only exception was the parameter for present bias for losses, with those who chose one pay-out having stronger present bias for losses, $t(160) = -2.02, p = 0.04$.

Next, it was explored if those who chose one of the most prominent tailored preference patterns differed on the obtained economic and psychological measures. We found no such differences for respondents pre-committing with one certain pay-out (*t* tests, all p 's > 0.12). A similar lack of evidence is observed for those who chose to commit with certain weekly pay-outs, constant (*t* tests, all p 's > 0.12), or slightly increasing (*t* tests, all p 's > 0.12). Several exceptions were observed: i) those pre-committing with weekly constant pay-outs discounted losses at a lower rate, $t(160) = -2.02, p = 0.04$, and ii) those pre-committing with weekly, slightly increasing pay-outs had more concave utility curvature for gains, discounted both gains and losses to a smaller extent and had less pronounced present bias for losses (*t* tests, all p 's < 0.03).

Finally, this lack of systematic association between the obtained measures and tailored incentives was confirmed by a series multiple linear or logistic regression analyses, in which subject characteristics, economic preferences and psychological traits were included stepwise as predictors for each tailored incentive dimension (for model specifications used and regression results see the Online Supplements of this dissertation).

Table 4.3. Descriptive statistics for demographic variables and psychological traits measures

Health behaviors	<i>n</i>	<i>M</i>	<i>SD</i>	Psychological measures	<i>n</i>	<i>M</i>	<i>SD</i>
Cigarettes (per week)	182	1.05	2.64	Trait self-control	163	3.13	0.58
BMI	182	21.85	4.24	Cognitive reflection	147	1.65	1.16
Alcohol (glasses/week)	182	7.91	9.40	EPQ - Neuroticism	136	0.58	0.20
Exercise (days/week)	182	2.85	1.67	EPQ - Extraversion	136	0.50	0.18
				EPQ - Psychoticism	136	0.43	0.19
				EPQ - Social desirability	136	0.52	0.23

Table 4.4. Descriptive statistics for tailored incentives selection using newly developed tool

Dimension	Options:	Count (%)
Pre-commitment (n = 182)	No	59 (32%)
	Yes	123 (68%)
Timing (n = 182)	One pay-out	85 (47%)
	Weekly pay-out	97 (53%)
Sequence (n = 97)	Strongly increasing	14 (14%)
	Increasing	37 (38%)
	Constant	31 (32%)
	Decreasing	11 (11%)
	Strongly decreasing	4 (4%)
Risk (n = 182)	1% (highest risk)	3 (2%)
	2% - 9%	0 (0%)
	10% - 39%	6 (3%)
	40% - 69%	40 (22%)
	70 - 99%	46 (25%)
	100% (no risk)	87 (48%)

Discussion

Heterogeneity in preferences for financial incentives for health behavior change has rarely been studied (one of the few examples being: Halpern et al., 2015), and, thus, it is unclear who responds to financial incentives and why (Paloyo et al., 2015). To provide policymakers with some support in the choice between different (behaviorally inspired) financial incentive schemes, this study explored the preferences of respondents themselves. More specifically, this study aimed to explore heterogeneity in the type of financial incentives individuals prefer and if the behavioral insights oft used in practice to motivate the choice for a particular design are associated with these preferences.

Surprisingly, the findings of this study indicate a large majority students would commit their own money to reach their exercise goals, whereas the work by Halpern and colleagues (2015) suggested uptake of such deposit contracts to be much lower. Furthermore, even though lottery incentives with small chances of receiving a relatively large sum have been used successfully (e.g. Haisley et al., 2012, Kimmel et al., 2012, van der Swaluw et al., 2018, Volpp et al., 2008), very few students selected incentives with low chances (<1% - 5%) of winning a prize for themselves. These tailored preferences were not systematically related to any of the behavioral insights often used to motivate implementation of behaviorally inspired incentives in practice (or to any of the measured health behaviors and psychological measures). Hence, although autonomy is likely increased by allowing individuals full freedom to design their own financial incentives using a tool like the one developed for this study, the results reported here provide no insight into why individuals prefer particular incentive schemes and if this will improve cost-effectiveness. Before providing interpretations based on this null result, and discuss explanations for the lack of evidence, several methodological limitations deserve noting.

First, the preferences reported here are obtained from students, and may not apply to the populations in which financial incentives are used to promote health behavior, such as individuals motivated to change their behavior (e.g. Halpern et al., 2015, van der Swaluw et al., 2018), or people in lower/middle income countries (for a review, see Ranganathan and Lagarde, 2012). For example, census data show that the young and highly educated exercise more than other populations (CBS/RIVM, 2018), and students may thus need fewer incentives to go to the gym twice a week. Second, all preferences obtained in this study were for hypothetical outcomes. In other words, this study investigated the association between hypothetical financial incentives for exercise and economic preferences elicited over hypothetical monetary outcomes. Although earlier work has suggested that preferences for hypothetical and real outcomes are not qualitatively different (Camerer and Hogarth, 1999, Hertwig and Ortmann, 2001), generally the use of real outcomes is preferred in behavioral experiments in health, as hypothetical incentives may lead to increased measurement error (Galizzi and Wiesen, 2018). Third, the experimental set-up and instructions used for this study could have had an influence on the findings reported in this study. For example, students were instructed to tailor incentives for going to the gym twice in order to reach a weight loss goal, and also explicitly told that they would know which incentives would motivate them. However, no further information is provided on their weight loss goal, the nature of the activities they should (imagine themselves to) perform in the gym, or how they should know what motivates them. As such, the instructions could have been open to alternative interpretations, which future work could remedy by using different instructions and focus on individuals' own health promotion goals. Furthermore, all measures obtained in this study were filled in by respondents only after they reported their preferences for tailored incentives. Without any counterbalancing procedures this study could not be controlled for ordering effects, as for example found in Carlsson and colleagues (2012). However, economic preferences were generally in line with those found in earlier applications of the methods used in this study (Abdellaoui et al., 2016, Bruhin et al., 2010, and Chapter 7 of this dissertation), and no association was observed between these preferences and the incentives selected. Hence, it is unlikely that respondents aimed to be consistent between the two parts of the experiment.

This study reports an exploration of the economic preferences that influence the incentives individuals prefer and found none to be systematically associated with self-selected incentives. This null result can mean one of two things: i) no such association exists or ii) the methods used failed to capture this association between economic preferences and (tailored) incentive. One explanation for the former, as suggested by Halpern and colleagues (2015) for the low uptake of deposit contracts (which were the most effective incentive design in their study), respondents may lack the sophistication to select financial incentives that would benefit them the most (e.g. they have insufficient knowledge of their own preferences, as found by: Hey and Lotito, 2009). This would explain why no association could be found between behavioral insights such as loss aversion and probability weighting, and the incentive dimensions these constructs are hypothesized to amplify.

On the other hand, the null result reported in this study may also be explained by a lack of external validity of economic preferences or insufficient statistical power to detect small but relevant effects. For example, earlier work has questioned whether the elicitation of economic preferences has bearing on decision-making in the field at all (Galizzi et al., 2016a, Galizzi and Navarro-Martínez, 2018, Schram, 2005). As such, one could question the usefulness of measuring economic preferences in the context of provision of financial incentives. Nonetheless, Björkman Nyqvist et al. (2018) did find a strong association between risk preferences and lottery incentives for uptake of secondary prevention in a field study in Lesotho. Compared to the field study by Björkman Nyqvist et al. (2018), this study used hypothetical incentives and a relatively small sample. Hence, the smaller statistical power and possibly increased noise related to hypothetical incentives may explain why this large effect did not extend to the lab.

To conclude, this study has several implications for future research and policy. The descriptive results reported suggest that that preferences for financial incentives differ between individuals. Hence, governments or organizations aiming to use financial incentives could, for example, use this tool or a similar one to study these preferences in their target population as input for the design of their interventions. Furthermore, it could be studied if tailoring incentives improves their cost-effectiveness, for example because tailoring increases motivation through enhanced autonomy, or because a subgroup of sophisticated individuals select incentives that are especially beneficial to them. An alternative way forward, to be explored either in future research or policy, is to assign individuals financial incentives that fit their economic preferences. However, although behavioral insights are oft used to motivate one-size-fits-all behaviorally inspired financial incentives, the theoretical or empirical basis for assigning individual-level tailored incentives is currently lacking.

5

Trust me; I know what I
am doing. Does domain
experience reduce
preference reversals in
decision making for others?

Abstract: When comparing methods to elicit people's preferences, preference reversals may be observed, i.e. choosing A over B when preferences are elicited using one method, while preferring B over A using another. This poses a significant problem for theoretical and applied research. We used a sample of medical and economics students to investigate preference reversals in the health and financial domain when choosing for patients/clients. We explored whether preference reversals are associated with domain experience and tested whether using guided 'choice list' elicitation reduces reversals. Our findings suggest that preference reversals were more likely to occur for medical students, within the health domain, and for open-ended valuation questions. Familiarity with a domain reduced the likelihood of preference reversals in that domain. Although preference reversals occur less frequently within specialist domains, they remain a significant theoretical and practical problem. The use of clearer valuation procedures offers a promising approach to reduce preference reversals.

Introduction

The elicitation of preferences, i.e. finding out if one prefers A over B or vice versa, is central in economics, and as such relevant to many topics studied in health economics, such as health state valuations (Torrance, 1976), multi-criterion decision analysis (Baltussen and Niessen, 2006), patient preferences (Ryan et al., 2001), and studies on physician behavior (Hennig-Schmidt et al., 2011). Many different methods are used to elicit preferences in the relevant target group, including well-known methods like willingness to pay (Himmler et al., 2020), time trade-off (e.g. Dolan et al., 1996), and discrete choice experiments (e.g. Green and Gerard, 2009). A disturbing finding is that using different methods; different preference orderings may be obtained. This phenomenon is typically referred to as *preference reversals*, i.e. the situation in which a person prefers A over B using one method, but B over A when using another method. This raises fundamental questions regarding the stability of preferences, preference elicitation, and the methods used in that context. Typically, under the assumption of procedural invariance, i.e., that preferences are independent from elicitation procedures (Chapman and Sonnenberg, 2003), one would assume that any observed preference ordering would also be observed when using another (common) elicitation method or using different operationalisations of the same method (e.g. Attema and Brouwer, 2008).

To illustrate this, imagine a person who indicates that she prefers surgery over physiotherapy for a given condition in a discrete choice experiment. Given this observation, we would expect her also to be willing to pay more, or at least not less, for surgery than for physiotherapy if we would elicit her willingness to pay for both options (*ceteris paribus*) as well. If this is the case, her preferences could be classified as consistent, but in practice, her willingness to pay for physiotherapy could be higher than that for surgery. This is an example of a preference reversal, and such reversals have been shown to be relatively common in previous studies (Grether and Plott, 1979, Tversky and Thaler, 1990). Preference reversals appear to be a robust violation of procedural invariance, which typically occurs when comparing preferences for risky outcomes elicited using different methods (Seidl, 2002) or different operationalizations of the same method (Attema and Brouwer, 2008). In a classic example, Slovic and Lichtenstein (1971) offered subjects two risky lotteries, referred to as the P-bet and the \$-bet. The former included a high chance of a moderate reward (e.g. 95% chance of winning 40\$, or lose 10\$ otherwise), while the latter involved a lower chance of a high reward (e.g. 15% chance of winning 160\$ or lose 15\$ otherwise). Preferences were first elicited using *direct choice*, i.e. subjects were asked to indicate which lottery they would choose. Next, subjects were asked to indicate the monetary values they would assign to both lotteries, i.e. their *valuation*. Slovic and Lichtenstein (1971) found that for lotteries with similar expected values, subjects chose the P-bet over the \$-bet, but assigned a higher monetary value to the \$-bet compared to the P-bet. This finding has been replicated frequently (e.g. Hamm, 1979, Reilly, 1982, Seidl, 2002) and constitutes a preference reversal, as economic theory predicts that the preferred lottery should also have been assigned a higher valuation. Preference reversals have been documented extensively for monetary outcomes (e.g. Grether and Plott, 1979, Lichtenstein and Slovic, 1971, Oliver and Sunstein, 2019) as well as for health outcomes (e.g. Attema and Brouwer, 2013, Oliver, 2006, Oliver, 2013b, Stalmeier et al., 1997).

If preferences are no longer stable, but depend on and can reverse between different elicitation methods and procedures, it is no longer possible to determine which (if any) method yields ‘true’ preferences (Braga and Starmer, 2005). Hence, preference reversals offer substantial methodological challenges, but also form a general and fundamental problem to applied work and decision-making in health and other settings. By now, the existence of preference reversals has been established in multiple domains. However, improving our understanding of causes and potential ways to lower preference reversals in different contexts remains crucial. Moreover, previous studies focused on individual decision making, that is, making choices for oneself. Less is known about choices on behalf of others and the influence of experience and training on preference reversals. Tversky and Thaler (1990) discussed preference reversals from the perspective of a social planner, while others applied this frame (Baron and Ubel, 2001) by using the person trade-off method. To our knowledge, the only study that investigated preference reversals in decisions made on behalf of others is by Oliver (2013b). In this study, we focus exclusively on decision-making as an agent on behalf of others in two domains. Since medical students, and to a lesser degree economics students, are trained to make decisions about others’ financial assets or health, respectively, this paper investigates an entirely new aspect of the preference reversal phenomenon.

In this study, therefore, we perform an elaborate investigation of preference reversals focusing on several relevant and novel aspects. First, we study preference reversals in the context of choices in both the financial and health domain. This allows a direct comparison of the degree of preference reversals in both domains. Second, we study preference reversals in the context of choices on behalf of others. In many real-life situations, we rely on the advice and actions of trained agents who inform or advise on choices or are delegated to make these. Examples include financial experts selecting investment portfolios or physicians proposing preferred treatment options. Preference reversals in such ‘experts’ are highly relevant to study, but have received little attention (see Oliver, 2013b, for an exception). Third, given our experimental design and sample selection, we are able to study the influence of domain experience on preference reversals. One might expect that training in a particular area (financial or medical choices) could reduce preference reversals in choices in that domain and increase consistency. Finally, we investigate whether clearer preference elicitation techniques result in fewer preference reversals.

The paper is structured as follows; firstly, we will discuss the literature on preference reversals and attempts to reduce them in the background section. We then continue to explain our experimental procedure in the methods section and finish with presenting our results and discussing them in the context of the literature.

Background

As indicated, preference reversals in individual decisions for oneself have been studied extensively for monetary outcomes, using many different settings and methods (for a review, see: Seidl, 2002). Preference reversals in decisions related to health outcomes by now have been documented in several studies as well (Bleichrodt and Pinto Prades, 2009, Oliver, 2006), Oliver (2013b), (Pinto-Prades et al., 2018, Stalmeier et al., 1997). To our knowledge, the only study directly comparing preference reversals in choices regarding health and money is that of Oliver and Sunstein (2019), who found a higher rate of preference reversals for

health. However, their study used different samples for each outcome domain, implying that differences in sample composition may partly explain their findings. Our study is the first to allow a direct, within-subject comparison between rates of preference reversal for health and money. Furthermore, we add to the literature on preference reversals by testing whether their occurrence is affected by two experimental conditions that tie into two explanations for the preference reversal phenomenon.

First, we study whether relevant domain experience for the choice context affects rates of preference reversals. This is investigated by having groups of respondents who (selected into) *domain-relevant training*, through their choice to become a physician or economist, i.e. medical students and economics students. Note that besides domain-relevant training, these groups may also differ in terms of skills and traits that precede and affect self-selection into these educational tracks, for example, the wish to help others in medical students (e.g. Galizzi et al., 2015, Godager and Wiesen, 2013). The effects of domain experience and cultural background on risk perception have been studied before in the monetary domain, with mixed findings on the effect of domain experience. Bontempo et al. (1997) found no significant difference in risk ratings for investments between students and security analysts with the same cultural background. Fraser-Mackenzie et al. (2014) found that self-stated domain experience in gambling, in general, did not affect risk preferences, but that participants with more experience were more risk-seeking when risks were described in fractional odds. More recently, Chang et al. (2016) found that stock investors with more domain experience in financial decision making were less risk-averse in picking portfolios and showed more confidence in their financial decision making. Although preference reversals are usually observed for risky gambles, it is currently unclear if these mixed findings on the effect of domain experience also affect the rate of preference reversals.

It is well-known that preference elicitation (for risk) may contain noise or imprecision (Bhatia and Loomes, 2017), which may be more likely if preferences are elicited for outcomes that one has no decision experience with or interest in. According to Butler and Loomes (1988, 2007), indicating the value of a lottery (i.e. providing a certainty equivalent) is a difficult task which leads to imprecision, and this imprecision may explain part of the systematicity of preference reversals. Hence, the high rates of preference reversal observed in earlier studies on health outcomes (Bleichrodt and Pinto Prades, 2009, Oliver, 2006, Oliver, 2013b, Stalmeier et al., 1997), maybe in part explained by the fact that most samples in these studies are generally unfamiliar with decisions about health. Indeed, Beshears and colleagues (2008) indicate that a lack of experience and choice complexity increase the occurrence of decision-making errors in preference (such as preference reversals). Pinto-Prades et al. (2018) provided more support for the role of imprecision in producing preference reversals by showing how preference reversals for health outcomes can be reduced by repeating preference elicitations. This leads to a prediction that fewer preference reversals may occur for health outcomes when preferences are elicited from those with domain-relevant training and those who are faced with these decisions more frequently (experience) – i.e. when eliciting actual physicians' preferences. In behavioral health economics experiments, physicians' preferences are often studied by recruiting medical students (e.g. Hennig-Schmidt et al., 2011), as their preferences differ only slightly (Brosig-Koch et al., 2016). In this study, to test the effect of domain experience through domain-relevant training, we will compare the degree of preference reversals within decisions by medical students for health and money, to the degree of preference reversal observed for economics students in these two domains. If

experience and training would matter, one might expect less preference reversals in the health domain for medical students and in the money domain for economics students.

Second, preference reversals are often explained by overpricing of the \$-Bet (i.e. low chance to gain a high outcome) as a result of scale compatibility (Tversky et al., 1990). This hypothesis suggests that people focus on different aspects of lotteries depending on the elicitation method. In direct choice, they give more attention to probabilities, which benefits the P-Bet (i.e. the high chance of winning a moderate amount), as this bet has a higher chance of yielding a positive result. In valuation, which is typically operationalized by using open-ended questions (e.g. “For what price would you sell this lottery?”), subjects focus on the unit in which they should express their valuation. In the study by Tversky and colleagues (1990) this focus on monetary amounts favours the \$-Bet and therefore could explain the relatively high rates of preference reversals. There is substantial evidence both for health (Attema and Brouwer, 2013) and money (Bateman et al., 2007a, Bostic et al., 1990) showing that preference reversals are more pronounced when comparing choice-based and open-ended valuation, as opposed to when exclusively choice-based methods are used. Furthermore, Oliver (2013b) argued that people are unlikely to have fixed preferences for unfamiliar goods and may use unstable heuristics when asked to value them using open valuation. As a result, there have been attempts to simplify open-ended valuation elicitation for respondents. For example, Oliver (2013b) tried an assisted valuation procedure by presenting respondents a selection of amounts to pay for a risky operation but found no notable differences to open valuation. Therefore, it remains relevant to investigate whether preference reversals can be reduced by using a preference elicitation method that is exclusively choice-based because scale compatibility may be less pronounced in that case, as opposed to open-ended valuation.

In this study, we will apply a choice-based method that has become increasingly popular in experimental economics (since Holt and Laury, 2002) due to its simplicity: the ‘choice list’. These choice lists are often used to quickly and reliably measure risk and time preferences (Andersen et al., 2006, Andersen et al., 2008), and allow to infer the value assigned to bets from a series consecutive choices presented vertically aligned on a single page. We test if the use of computer-assisted choice lists (see Figure 5.2, for example), as opposed to open-ended valuation, reduces the occurrence of preference reversals in decision-making for both health and money.

Methods

Sample and experimental design

To ensure that every participant had at least some prior experience with choices in one of the domains, we aimed to only recruit economics, business and medical students beyond their first year of studies. Our full sample of 252 students comprised of 129 medical students, 121 business and economics students (henceforth: economics students) and two other students (removed from the sample). Recruitment of these students differed depending on their discipline. Economics and business students were recruited from the subject pool of the experimental laboratory at Erasmus School of Economics, while medical students were recruited through messages in the virtual learning environment of two University Medical Centres (in Rotterdam and Leiden). Subjects were paid a flat fee of 10 Euros (paid out as a gift voucher) for participating in the experiment.

Both groups of students completed an online experiment, which was operationalized in Qualtrics Survey Software, with a two by two within-subjects factorial design, using the following two factors: i) outcome domain (health vs financial), and ii) valuation procedure (open-ended vs choice list)¹⁸. This design allows us to study preference reversals within-subjects in four blocks and allows between-subjects comparisons based on discipline (i.e. economics or medicine). An overview of our experimental design is provided in Figure 5.1. To avoid ordering and learning affects the order of outcome domains and valuation procedures were randomized.

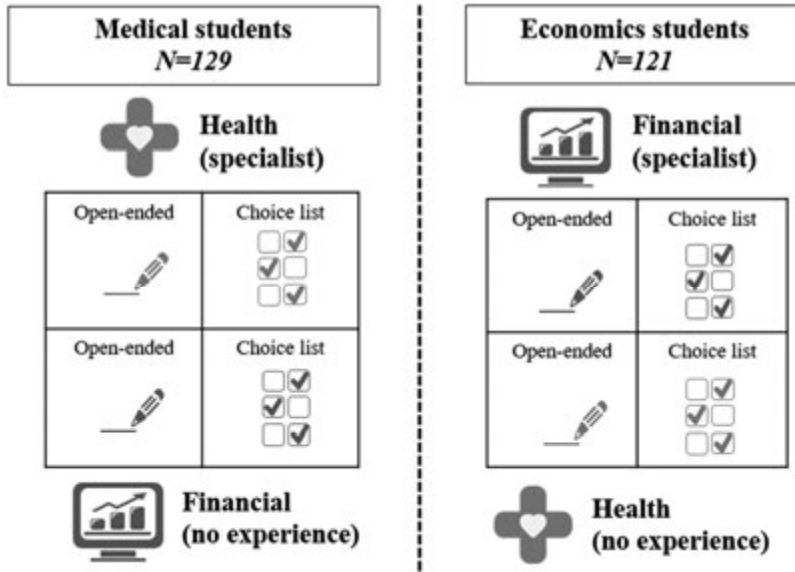


Figure 5.1. Survey design of the two domains and valuation procedures

Experimental procedure





Several screening questions were in place, to avoid recruiting first-year students and students from other disciplines. The online experiment started with general instructions and a practice block. Afterwards, participants completed a total of 12 questions eliciting their preferences for health and investment decisions (on behalf of others) with one choice and two valuation questions for each condition. Both scenarios began with an introduction page informing participants which role they would have in the experiment that followed. Graphical elements were added to inform respondents which type of question they were answering and to reduce the repetitiveness of the questions. After completing the 12 questions, demographics were collected. More specifically, we collected information on age, gender, statistical competency, and year of study.





¹⁸ We also piloted a condition aimed at reducing preference reversals by using natural frequencies to communicate risks, but due to a programming error this data could not be included.

Eliciting preference reversals

The questions per condition all followed a similar structure, following the classic study by Slovic and Lichtenstein (1971): i) a strict choice between two risky lotteries with similar expected values (henceforth P-bet and \$-bet), ii) valuation of P-bet, iii) valuation of \$-bet (for an overview of P-bets and \$-bets used, see Table 5.1). The order of these three questions was randomized within each condition. We recorded a preference reversal if a respondent chose the P-bet over the \$-bet in the direct choice, but at the same time valued the \$-bet strictly higher in the valuation questions. This commonly observed reversal pattern is usually referred to as a ‘predicted preference reversal’ as it is what is predicted by scale compatibility (Tversky et al., 1990). Preferring the \$-bet while assigning a strictly larger value to the P-bet is defined as an ‘unpredicted preference reversal’. We will interpret subjects indicating to prefer one bet in direct choice while assigning it a higher or equal value in valuation as having consistent preferences (as this combination of preferences does not violate procedural invariance).

Table 5.1. P-bets and \$-bets used for health and financial outcomes in all four conditions

 Health Open-ended 		 Health Choice list 	
P-bet	\$-bet	P-bet	\$-bet
80% of 5 years	55% of 9 years	90% of 7 years	50% of 14 years
15 % of 0 years	20% of 0 years	5 % of 0 years	35% of 0 years
5 % of -1 years	25% of -1 years	5 % of -2 years	15% of -2 years

 Financial Open-ended 		 Financial Choice list 	
P-bet	\$-bet	P-bet	\$-bet
85% of 1600€	35% of 5500€	95% of 2000€	33% of 7000€
10% of 1000€	15% of 2700€	4% of 500€	33% of 1350€
5% of -550€	50% of -1800€	1% of -1500€	34% of -2500€

Operationalization of outcome domains (health vs financial)

Each condition began with an instruction that set the stage for respondents. In both domains, respondents hypothetically advised a person on a decision between two risky prospects. In the financial scenario, respondents advised clients on how to invest their money in different portfolios. The health scenario involved recommending treatment options for a terminally ill patient, where patient health status was described by using the dimensions of the EQ-5D instrument. Whereas in the original set-up by Slovic and Lichtenstein (1971), which was extended to health outcomes by Oliver (2006, 2013b), risky prospect were two outcome mixed gambles (comprising of a gain and a loss), Table 5.1 shows that the P-bets and \$-bets

in this study used three outcomes. The third outcome was included to increase realism¹⁹, as both investment and medical decisions typically have at least three outcomes: a gain (high return on investment or medical treatment is successful), ‘the status quo’ (moderate return on investment or medical treatment is unsuccessful), and a loss (portfolio value decreases or side-effects of medical treatments). In each question, graphical elements like those in Figure 5.1 were used to emphasize the outcome domain and valuation procedure being used, which may have increased subjects’ attention to changes between conditions. We can test the effect of domain-relevant training on preference reversals, as our sample of medical students should have more experience with decisions about health compared to economics students, while economics students should have more experience with financial decision-making, respectively.

Operationalization of valuation procedure (open-ended vs choice list)

Open-ended valuation is the typical procedure in papers studying preference reversals (Attema and Brouwer, 2013, Grether and Plott, 1979, Lichtenstein and Slovic, 1971, Oliver, 2013b). For health outcomes, open-ended valuation was operationalised as follows: students were instructed to compare the P-bet (\$-bet) to a treatment yielding some amount of life years in perfect health for certain, where students were asked to provide the minimum amount of life years that would lead them to recommend patients to take this certain treatment over the P-bet (\$-bet). For financial outcomes, the open-ended valuation was operationalized as follows: students were asked to compare the P-bet (\$-bet) to a government bond yielding a sure gain and asked to indicate how large this gain should be for the bond to be equally good to the P-bet (\$-bet). In both outcome domains, students were required to provide this certain amount of life years or money in an open answer field, i.e., students reported a certainty equivalent. Choice list valuation was operationalized by simplifying these valuation questions by offering respondents a list of increasing amounts of money (in increments of 1,000\$, followed by 100\$ increments) or life years (in yearly increments) to choose from. Figure 5.2 shows an example of such a choice list valuation procedure for valuation of a P-bet, where at some point students switch from preferring the P-bet to a certain outcome. As is usual in choice list methodology (Holt and Laury, 2002), the certainty equivalent is obtained by taking the average of the certain outcome above and below the switching point (see Figure 5.2 for examples). This procedure was guided as the choice lists were programmed to prohibit multiple switching points and choices that violated dominance.

¹⁹ To check the realism of our P-bets (\$-bets) and the instructions used for medical decision-making, we consulted a physician. Some minor changes were made to the framing (e.g. we increased the age of the patient whom students are to imagine they would be advising).

<u>Financial example</u>				<u>Health example</u>			
<i>\$-bet</i>				<i>P-bet</i>			
33% of 7000\$				90% of 7 years			
33% of 1350\$				5 % of 0 years			
34% of -2500\$				5 % of -2 years			
	A	B		A	B		
Government bond yielding \$0	<input checked="" type="checkbox"/>	<input type="checkbox"/>	\$-bet	P-bet	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Living healthily for 0 additional years
Government bond yielding \$1000	<input checked="" type="checkbox"/>	<input type="checkbox"/>	\$-bet	P-bet	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Living healthily for 1 additional years
Government bond yielding \$2000	<input checked="" type="checkbox"/>	<input type="checkbox"/>	\$-bet	P-bet	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Living healthily for 2 additional years
Government bond yielding \$3000	<input checked="" type="checkbox"/>	<input type="checkbox"/>	\$-bet	P-bet	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Living healthily for 3 additional years
Government bond yielding \$4000	<input checked="" type="checkbox"/>	<input type="checkbox"/>	\$-bet	P-bet	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Living healthily for 4 additional years
Government bond yielding \$5000	<input type="checkbox"/>	<input checked="" type="checkbox"/>	\$-bet	P-bet	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Living healthily for 5 additional years
Government bond yielding \$6000	<input type="checkbox"/>	<input checked="" type="checkbox"/>	\$-bet	P-bet	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Living healthily for 6 additional years
Government bond yielding \$7000	<input type="checkbox"/>	<input checked="" type="checkbox"/>	\$-bet	P-bet	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Living healthily for 7 additional years

Figure 5.2. Hypothetical response to choice list valuation of a \$-bet (financial) and P-bet (health), yielding certainty equivalents of 4,500\$ and 3.5 years, respectively.

Data analysis

Preference reversals for each scenario were first analysed descriptively by creating a dummy variable, which indicated if a preference reversal occurred or not. Subsequently, we compared preference reversals by study discipline, outcome domain and valuation procedure using Chi-squared tests. Next, we analysed our findings by employing a logistic regression (in R using *lmerTest*) with random subject effects, which predicted preference reversals by the following fixed effects: a) domain (financial vs health), b) discipline (economics vs medical students), c) procedure (choice list vs open-ended valuation, d) domain-relevant training (domain x discipline interaction) and e) interaction term for procedure and discipline. This analysis allowed us to determine to what extent the chance of showing a preference reversal was affected by our experimental conditions.

Results

In addition to the two students without domain-relevant training in the areas we were investigating, two students were excluded who reported being in their first year of studies. Accordingly, the final analysis is based on a sample of 248 respondents.

Descriptive statistics

Sample characteristics for these two groups of students can be found in Table 5.2. Comparisons between the two groups yielded some significant differences, showing that economics students (relative to medical students) were more likely to be male (Chi-squared test, $p < 0.002$), and reported being in a higher study year and more competent in statistics (Wilcoxon tests, p 's < 0.02). Table 5.3 shows the overall results of this online experiment, which indicate that predicted preference reversals were the most occurring combination of preferences in all conditions. Furthermore, only very few unpredicted preference reversal occurred, representing just over 1% of all combinations of preferences. Hence, we will study both reversals combined, and for brevity refer to these as '*the rate of preference reversal*'.

Table 5.2. Sample characteristics by study discipline

	Economics (n=119)		Medicine (n=129)		Total (n=248)	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Age	21.60	1.94	21.43	2.24	21.51	2.10
Stat. comp. ²⁰	2.94	1.02	2.51	0.82	2.72	0.94
Study year	3.81	1.32	3.58	1.69	3.69	1.53
Gender	Female	Male	Female	Male	Female	Male
	58	61	104	25	162	86

Comparisons by students' discipline, outcome domain, and valuation procedure




When we sum preference reversals (i.e. predicted and unpredicted), we find that combined for all conditions economics students (44.7% reversals) revealed a significantly lower (Chi-squared, $p < 0.001$) rate of preference reversals than medical students (56.6% reversals). Furthermore, preference reversals occurred more frequently for outcomes in the health domain (54.4%) compared to the financial domain overall (47.6%), this difference was significant (Chi-squared, $p = 0.03$). Finally, we find that that the rate of preference reversals was significantly higher (57.4% reversals) for open-ended valuation (Chi-squared, $p < 0.001$), compared to choice list valuation (44.6 % reversals).




When comparing individual conditions, an effect of domain experience appeared to occur. Medical students had a 6.3 percentage point (pp) lower rate of preference reversals in their

²⁰ 1 indicating "I had no statistical training", 2 "I feel somewhat competent with statistics", 3 "I know my way around statistics, but I'm not an expert", 4 "I feel competent in statistics", 5 "My specialization is statistics".

area of expertise (59.3% reversals for health outcomes) when using open valuation as compared to financial outcomes (65.6% reversals), but not statistically significant (Chi-squared, $p = 0.30$). Economics students, on the other hand, had a significant 14.6 pp difference (Chi-squared, $p < 0.05$) between financial (44.7% reversals) and health (59.3% reversals) outcomes and were as expected more consistent in the financial domain (their area of expertise). By using choice list valuation, both economics and medicine students were more consistent compared to open valuation (i.e., lower rates of preference reversal). The most substantial reductions in the rate of preference reversals through choice lists could be observed outside of the respondent’s area of expertise. The rate of preference reversals (choice lists 43.1% vs. 59.3% open valuation) of economics students using choice lists was 16.2pp lower (Chi-squared, $p < 0.05$) in the medical domain as opposed to a 11.4 pp (Chi-squared, $p < 0.10$) reduction (33.3% vs. 44.7%) in the financial domain. Medical students showed a not significant 3.9 pp (Chi-squared, $p = 0.53$) reduction in the rate of preference reversals in the health domain (55.5% vs 59.4%) and significant a 19.5 pp (Chi-squared, $p < 0.05$) reduction in the financial domain (65.6 vs 46.1) when preferences were elicited with choice lists. This suggests a trend that choice lists help to reduce preference reversals, especially so in unfamiliar domains.

Table 5.3. Overall frequency distribution for combinations of preferences per condition

 Health Open-ended 		 Health Choice list <input type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>		Interpretation Predicted reversal Unpredicted reversal Consistent Consistent
Pattern	Observation (%)	Pattern	Observation (%)	
PS	147 (59.3%)	PS	120 (48.4%)	
SP	0 (0.0%)	SP	3 (1.2%)	
PP	77 (31.0%)	PP	89 (35.9%)	
SS	24 (9.7%)	SS	36 (14.5%)	

 Financial Open-ended 		 Financial Choice list <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/>		Interpretation Predicted reversal Unpredicted reversal Consistent Consistent
Pattern	Observation (%)	Pattern	Observation (%)	
PS	137 (55.2%)	PS	94 (37.9%)	
SP	1 (0.4%)	SP	3 (1.2%)	
PP	82 (33.1%)	PP	85 (34.3%)	
SS	28 (11.3%)	SS	66 (26.6%)	

Note: The pattern P\$ indicates that the P-bet was chosen in the choice task, but that the \$-bet was valued strictly higher in the valuation task, while \$P indicates the reverse pattern. PP and \$\$ indicate a choice for a bet that was valued at least as good or higher (i.e., no violation of procedural invariance).

To substantiate our descriptive findings further, we ran a logistic mixed-effects regression, which allowed us to determine to what extent the chance of observing a preference reversal was affected by our experimental conditions. Table 5.4 shows several main effects. Preference reversals appear to be more likely to occur a) in the health domain, b) for decisions by medicine students, and c) for open valuations (as opposed to choice list elicitation). Furthermore, we observed a marginally significant interaction between discipline and domain (i.e., the effect of domain experience); medical students were less likely to show preference reversals in their ‘own domain’. Importantly, when exploring the robustness of our findings, we found that our main findings were mostly unaffected by including several demographics sequentially (e.g., statistical competency, sex, age, and year of study), even when we included interaction terms. The results of these analyses are available from the authors on request.

Table 5.4. Results of logistic mixed-effects regression predicting the preference reversal by our experimental conditions

	<i>Estimate</i>	<i>SE</i>	<i>Z</i>	<i>p</i>
Constant	-0.65	0.15	-4.22	<0.001
Main effects				
Discipline (medical)	0.79	0.25	3.18	0.001
Domain (health)	0.55	0.19	2.85	0.004
Procedure (open ended)	0.62	0.19	3.22	0.001
Interaction effects				
Domain experience (medical x health)	-0.49	0.27	-1.79	<i>0.07</i>
Discipline (medical) x Procedure (open)	-0.06	0.27	-0.25	0.80

Note: bold-faced p-values are significant at $p < 0.01$, italicized p-values are significant at $p < 0.10$.

Discussion

This study investigated whether domain experience and guided choice list valuation procedures reduced the rate of preference reversal in decision making for others for both health and money. Given that we studied preference reversals for both health and money, the results of this study can be compared to the extant literature in these two domains. Overall, we find preference reversals to occur frequently (occurring in 32 to 66% of the sample depending on the condition). These high rates of (predicted) preference reversals are in accordance with earlier studies for money (Grether and Plott, 1979, Lichtenstein and Slovic, 1971) and health (Oliver, 2006, Oliver, 2013b, Oliver and Sunstein, 2019, Stalmeier et al.,

1997). Some studies, often with designs that deviate more from the original set-up used by Lichtenstein and Slovic (1971), find somewhat lower rates of preference reversals – especially for health (e.g. Bleichrodt and Pinto Prades, 2009, Pinto-Prades et al., 2018). To our knowledge, only a single study exists which studied preference reversals for both money and health (Oliver and Sunstein, 2019). These authors compared preference reversals for health and money (and other domains) using different samples for each domain and found higher overall rates of preference reversal for health, which we confirmed in our study with direct within-subjects comparisons. Furthermore, for 3 out of 4 between-subjects comparisons, preference reversals occurred more frequently for health. Our study also allowed testing whether domain experience gathered by receiving domain-relevant training influenced preference reversals, by comparing the differences in rates of preference reversals between health and money for economics and medicine students. Note that this inevitably also captured potential effects related to self-selection into these educational programs. While we find a higher rate of preference reversal for medical students overall, we find a trend suggesting that the increase in rates of preference reversals from money to health is smaller for medical students (as shown by the regression results in Table 5.4). For example, when medical students completed the open-ended valuation, we found fewer preference reversals for health than for financial outcomes, suggesting that medical students were more consistent in their own specialist domain.

Furthermore, our design allowed comparing the effect of choice-based and open-ended valuations for both money and health. We found that choice-based valuations, using guided choice list elicitation, reduced the rate of preference reversals for both health and money, regardless of familiarity with the outcome domain. Hence, our findings confirm earlier work for health (Attema and Brouwer, 2013) and money (Bateman et al., 2007a, Bostic et al., 1990). New to our approach was the comparison of open-ended valuations and computer-assisted choice lists. The latter has only recently been introduced in preference elicitation in health economics (e.g. Arrieta et al., 2017, Attema and Lipman, 2018, Galizzi et al., 2016c, Irvine et al., 2019, Pinto-Prades et al., 2018). Given that we found a more substantial effect of valuation procedures as opposed to domain experience, may suggest that in our study²¹ scale compatibility (Tversky et al., 1990) played a larger role in generating preference reversals than imprecise preferences (Butler and Loomes, 2007). Furthermore, while this study allowed us to test if the consistency in choices is affected by the elicitation procedure and the familiarity with the outcome domain, we have no way of determining what the ‘true preferences’ of participants would be. Moreover, we cannot assert that observing fewer preference reversals implies that elicited preferences are more aligned with such ‘true preferences’.

Regardless of our attempts to reduce them, preference reversals remained prevalent. Earlier work provides several explanations for these findings. First, as has been shown by Pinto-Prades et al. (2018), the choice list methodology is a transparent and straightforward way to elicit preferences. This explicit transparency may have allowed subjects to deduce that the goal of this task was to observe an indifference between two outcomes. If respondents are aware of the goal of the task, this could lead to strategic choices or influences from previous choices. Other methods, e.g. the hidden choice-based procedure developed by Fischer and

²¹ Note that this study was not designed to compare the role of scale compatibility and imprecision.

colleagues (1999), reduce these influences by spreading elicitation over multiple items that occur in random order, and they have been shown to reduce that the rate of preference reversals (Fischer et al., 1999, Pinto-Prades et al., 2018). As such, seeing as choice lists are highly transparent, their use in this experiment might explain the high prevalence of preference reversals. Second, we opted to study preference reversals in decisions for others, as this is relevant in real life and in the context of economics and medicine students' training. Oliver (2013b) found that preference reversals occur more frequently in the social decision making contexts. In our experiment respondents advise others on decisions, and hence, one might object to us referring to these choices on behalf of others as 'preferences' (and inconsistencies as preference reversals). However, similar to Oliver (2013b), we decided to also use the established term preference reversal in a decision making for others context since the phenomenon is well established under this term in the literature.

Third, this experiment was completed using online survey software. Although several studies found little differences between lab and online studies (Birnbbaum, 2000, Dandurand et al., 2008, Germine et al., 2012, Riva et al., 2003), other studies found that completing research in online environments may lead to higher variances or more noise (e.g. von Gaudecker et al., 2008). In our study, more noise would have been reflected in higher rates of preference reversals, both predicted and unpredicted. Given that the number of unpredicted preference reversals was negligible (less than 1.5%), our results give little indication to expect a large effect of noise related to the online nature of the experiment. Fourth, we opted to reward subjects with a flat fee for participation, while generally providing incentives compatible with stated preferences is preferred in experimental and behavioral health economics (Galizzi and Wiesen, 2018). However, given that we aimed to study preference reversals for both money and health, such incentive-compatibility was difficult to implement and could lead to unwarranted differences between the two domains. Hence, in order to prevent apparent procedural differences between health and money, preferences were elicited with hypothetical and relatively large stakes. Finally, our sample comprised students of economics and medicine, which raises the question whether our findings generalize to i) the general public, ii) other trained professionals and their respective domains, and iii) actual medical professionals or economists. Given the main dimensions on which our sample differed from the general public (e.g., age, education level and wealth) which are related to risk attitudes (Halek and Eisenhauer, 2001, Hartog et al., 2002), investigating the effects of choice based elicitation in a general public sample would be an interesting venue for future research. The same goes for similar studies in other domains and 'expert samples'. Furthermore, although recruitment may be time-consuming, to further study the effect of domain experience on preference reversal, future work could recruit respondents working as trained experts in these fields, such as investment bankers (as in Abdellaoui et al., 2013a) or physicians (as in Brosig-Koch et al., 2016). Although these studies give no indication to expect qualitatively different decision-making, such future work could explore if the positive trend related to domain-relevant training is amplified when more decision experience is accumulated. Similar to previous studies on preference reversals (Attema and Brouwer, 2013, Bateman et al., 2007, Oliver 2013a, Oliver 2013b) we opted to vary the outcomes and their probabilities between the questions. Another interesting area for future research might, therefore, be to study whether part of the reversals can be attributed to the use of different probabilities between elicitation procedures and outcome domains.

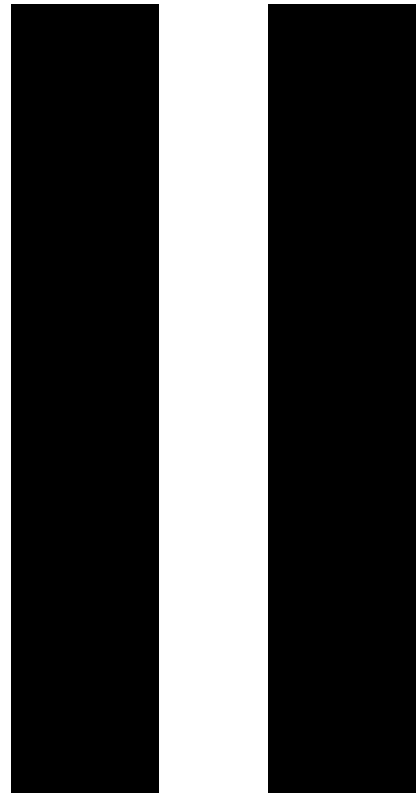
Conclusion

If observed preferences indeed depend on the way they are elicited, as we showed in this study, this is problematic. As long as revealed and stated preferences remain a cornerstone of research in health economics, such preference reversals offer a challenge to both empirical and theoretical work. For instance, when using ‘discovered preferences’ (Braga and Starmer, 2005) in practice, for example, when incorporating patient preferences in shared decision-making or health technology assessment, preference reversals pose fundamental questions. Whereas one may conclude from our work that preference reversals are robust, occur frequently and are especially prevalent in unfamiliar domains, we believe this study also may offer some guidance for preference elicitation in research and practice in the future. First, guided choice-based valuation, such as choice lists, may be a promising tool to obtain more consistent preferences. Second, although preference reversals were more common for decisions about health as opposed to financial decision-making, we find that medicine students show fewer reversals in their own domain. This effect could have several explanations, but a positive interpretation would be that a few years of domain experience could already improve consistency slightly. Hence, accumulated domain experience with the type of decisions for which preferences need to be determined (e.g. over years of medical practice) may help to be more consistent.



PART II

Applying behavioral insights to health state valuation



6

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

Chapter based on:

*Lipman, S.A., Brouwer, W.B.F. & Attema, A.E. (2020). What is it going to be,
TTO or SG? A direct test of the validity of health state valuation.*

Health Economics

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

Introduction

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

Both methods involve direct choices between two options, of which one is living in some health state Q for Y years. In TTO, respondents are offered an alternative: to live in perfect health (described as a state without health problems) for a shorter time, i.e. X years ($X < Y$). Respondents are asked to indicate duration X such that they are indifferent between both options. In practice, this indifference is evaluated by normalizing the utility of perfect health to 1 and that of being dead to 0, and assuming that the utility of duration is linear (Torrance, 1987). Under these strict assumptions, the obtained indifference reveals the TTO weight of state Q , which is obtained by: X/Y . SG offers a different alternative to living in health state Q for Y years: a lottery that results in perfect health for Y years (with probability p), or immediate death (with probability $1 - p$). Respondents are asked to indicate probability p such that they are indifferent between both options. In practice it is often assumed that respondents handle the risky lottery as modeled in expected utility theory. Under this strict assumption, probability p in the obtained indifference reveals the SG weight of state Q .

Bleichrodt (2002) proposed that the differences between TTO and SG weights could be explained by violations of the strict assumptions underlying the methods. SG responses are expected to be biased upwards (by probability weighting and loss aversion), while TTO responses are expected to be biased both upwards (by loss aversion) and downwards (by scale compatibility and utility curvature). Hence, Bleichrodt (2002) argued that i) the upward and downward bias in TTO might cancel out (to some extent) and ii) the difference between TTO and SG would diminish when these strict assumptions are dropped (e.g. as in prospect theory). While empirical evidence suggests that under less restrictive assumptions the differences between SG and TTO indeed diminish (van Osch et al., 2004, and also in Chapter 7 of this dissertation), the first claim by Bleichrodt (2000) is more difficult to address. If bias (partly) cancels out, this implies that that TTO weights are likely to be a better approximation of health state valuation. Testing the validity of QALY weights, however, would require a golden standard for 'true' health state preferences, to compare weights elicited with different methods to. It is safe to say that even the degree to which true preferences exist, and can be measured, is controversial (e.g. Braga and Starmer, 2005), let alone with what method they could be derived.

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

Methods

A total of 119 Business Administration students took part in this experiment and were rewarded course credit for participation. The sample consisted of 44 (37%) males and 75 (63%) females, with a mean age of 20 (SD = 0.99). The experiment took place in a university-based computer lab, in experimenter-led sessions of 30 minutes, run with up to four students. In the first and second part of this experiment (programmed in Shiny), after completing a practice task for a health state described as ‘chronic back pain’, subjects completed a block of TTO (SG) tasks for the following EQ-5D-5L health states: 21211, 31221, 31231, 31341, 33342 (see Table 6.1). Tasks (i.e. TTO or SG) and health states were presented in randomized order between-subjects. In the third part (i.e., the ‘validation’), subjects were presented with the implications of their responses in the first and second part (henceforth: implied QALY weights), and asked to validate them. After completing the experiment, subjects filled out a paper-and-pencil questionnaire measuring age, sex, and how difficult they felt the tasks were on a scale from 1 (not difficult at all) to 10 (very difficult)²².

Table 6.1. Health states used in this experiment including tariff elicited from Dutch value set for EQ-5D-5L (Versteegh et al., 2016)

Health state	Q1: 21211	Q2: 31221	Q3: 31231	Q4: 31341	Q5: 33342
Dutch Tariff	0.88	0.79	0.76	0.47	0.34
You have ... problems with walking	Slight	Moderate	Moderate	Moderate	Moderate
You have ... problems with washing and dressing yourself	No	No	No	No	Slight
You have ... problems with washing and dressing yourself	Slight	Slight	Slight	Moderate	Moderate
... pain or discomfort	No	Slight	Moderate	Severe	Severe
... anxious or depressed.	Not	Not	Not	Not	Not

TTO and SG task

Both methods were operationalized using the common 10-year duration (as is usual in valuation studies, see: Oppe et al., 2014), i.e. the period in the impaired health state was 10 years ($Y = 10$). TTO and SG indifferences were obtained through a bisection process with 5 choices. These 5 choices produced an indifference point, which subjects could confirm or change with a slider. These slider values were used to calculate the SG and TTO weights (as highlighted in the introduction). An overview of the instructions and supporting graphs used can be found in the Online Supplements of this dissertation.

²² We found no differences in task difficulty across TTO, SG and validation (paired Wilcoxon tests, p 's > 0.48).

Validation task

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

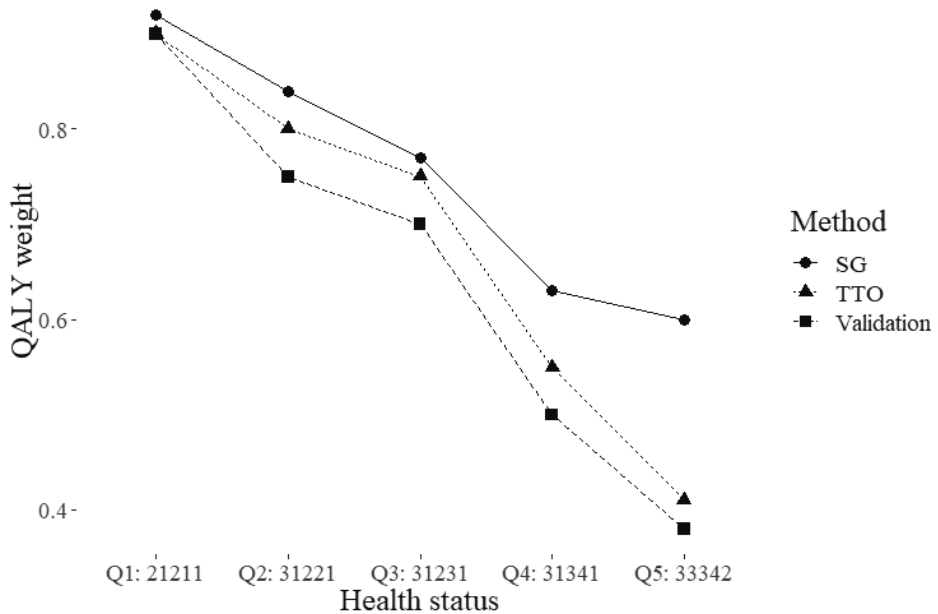


Figure 6.1. Median QALY weights elicited by TTO and SG, including the final QALY weights selected after validation

Results

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

Table 6.2. Overall respondent preferences for implied QALY weights and direction of change for confirmed QALY weights (n denotes the number of respondents, ↑ and ↓ denote upwards and downwards change) and divided by initial ordering of TTO and SG weights.

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

Note: ^a signifies that the proportion preferring this method's implied QALY weight (i.e. TTO or SG) is significantly higher than the other (Chi-squared test, $p < 0.05$), ^b signifies that (of those adjusting) a larger proportion adjusted QALY weights in this direction.

TTO and SG task

Figure 6.1 shows that health states received monotonically decreasing QALY weights for both TTO and SG (paired Wilcoxon tests, all p 's < 0.001). Inter-quartile ranges for TTO (SG) were between 0.21 and 0.40 (0.23-0.37) for all health states. The weights were only significantly different from the Dutch EQ-5D tariffs for Q4 and Q5 (Wilcoxon tests, p 's < 0.002). For all health states, TTO weights were significantly lower than SG weights (paired Wilcoxon tests, all p 's < 0.02). A within-subject comparison showed that for all health states, SG weights were most likely to be higher (55-66%). Sometimes, TTO values were higher

(22-34%) or both methods yielded equal values (6-13%). Chi-squared analyses showed that these counts were significantly different (all p 's < 0.001), i.e. the finding of higher SG values was also confirmed within-subjects.

Validation task: preferences for implied QALY weights

Table 6.2 shows the number of individuals that preferred the implied TTO or SG weights per health state. Overall, it can be concluded that for each health state (if a difference existed between TTO and SG), the majority of respondents indicated TTO weights represented the value of that health state best. Compiled for all health states, this finding was significant (Chi-squared test, $p < 0.001$). Analyses per health state showed that it was significant only for the two most severe health states. We also studied preferences for implied QALY weights within-subjects, showing that only 18% (5%) consistently preferred TTO (SG) values for all health states. Table 6.2 also shows preferences for implied QALY weights and adjustments split for which of the two weights was initially higher. Combined for all health states, these results show that if SG was higher than TTO, respondents were more likely to pick TTO (Chi squared test, $p < 0.001$). The opposite also holds. i.e. if TTO was higher than SG, SG weights was more likely to be preferred. Chi-squared test, $p < 0.001$). When analyzing separately for each health state, this holds for Q2 (only when TTO weights were smaller than SG weights) to Q5 (Chi squared tests, p 's < 0.05). If both weights were the same, no differences were observed in preferred implied QALY weight (Chi-squared test, all p 's > 0.16), i.e. the forced choice among two identical QALY weights was distributed independently.

Validation task: confirmed QALY weights

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

Discussion

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

We find that, according to respondents themselves, TTO weights are a better reflection of the value of a health state. This finding is in accordance with predictions by Bleichrodt (2002), who argued that upward and downward bias in this method may cancel out. Nonetheless, this cancelling out seemingly is not perfect, since respondents often adjusted their implied QALY weights. On average, the direction of this change was downwards, yielding lower confirmed QALY weights than implied by TTO. This may suggest that, as for example was found in Chapter 7 of this dissertation, the net effect of bias in TTO remains upwards. In this study, the magnitude of this remaining bias appears to increase with severity, with confirmed QALY weights being significantly lower than TTO weights for the two severest states only. Perhaps, directly correcting biases (as in Chapter 7 of this dissertation) could provide a further step towards valid QALY weights, rather than hoping biases in different directions cancel out.

A few limitations deserve noting. First, we used a relatively small student sample, which means that the generalizability of our findings may be questioned. We encourage future work to apply this approach in (larger) general public samples (for instance in the context of EQ-5D valuation). Second, we used our approach only in the context of health states better than dead, which might be remedied in future work using methods suitable for valuing health states worse than dead. Finally, an alternative explanation of our findings, which also reflects a fundamental difficulty in directly presenting and validating QALY weights, would be that our validation task has some resemblance to a visual analogue scale (VAS). It is well-known that VAS may suffer from other biases than TTO and SG do, and QALY weights elicited with VAS are generally lower than those elicited with TTO and SG (Bleichrodt and Johannesson, 1997, Robinson et al., 1997, Robinson et al., 2001). As any method of presenting QALY weights might suffer from biases, future work could, for example, test if alternative graphical or textual presentations lead to different conclusions.

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

7

QALYs without Bias? Non-parametric correction of time trade-off and standard gamble weights based on prospect theory

Chapter based on:

Lipman, S.A., Brouwer, W.B.F., & Attema, A.E. (2019). QALYs without bias? Nonparametric correction of time trade-off and standard gamble weights based on prospect theory. Health economics, 28(7), 843-854.

Abstract: Common health state valuation methodologies, such as standard gamble (SG) and time trade-off (TTO), typically produce different weights for identical health states. We attempt to alleviate these differences by correcting the confounding influences modeled in prospect theory (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 suggest new opportunities to account for bias in health state valuations, but also the need for further validation of resulting weights.

Introduction

In cost-utility analyses (CUA), incremental costs of medical technology are compared with incremental health benefits, commonly expressed in Quality Adjusted Life Years (QALYs). These QALYs (Pliskin et al., 1980) are obtained multiplying prospective life years by weights, sometimes referred to as ‘utilities’. QALY weights represent health-related quality of life, such that 0 represents the subjective weight of the state ‘dead’ and 1 that of full health. Several methods are used to obtain QALY weights, most notably Standard Gamble (SG) and Time Trade-Off (TTO). Empirical work, however, has demonstrated that QALY weights differ systematically between these two elicitation methods, with SG weights being higher than TTO weights (e.g. Bleichrodt and Johannesson, 1997, Torrance, 1976). As a consequence, QALY weights and, hence, outcomes of economic evaluations may depend on the health state valuation (HSV) method used.

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

PT is characterized by four tenets (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992). These are: (1) reference dependence – utility derived from a good is defined over differences from a reference point (RP), instead of over the overall consumption of that good; (2) loss aversion - the utility function has an inflection point at the RP and is steeper for losses than for gains; (3) diminishing sensitivity - utility 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 overweighs small probabilities and underweighs large probabilities (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992). PT is usually applied to decisions about money, but has also been extended to health outcomes (Bleichrodt and Pinto, 2000, Miyamoto and Eraker, 1989). Importantly, as Bleichrodt (2002) proposed, the tenets modelled in PT will likely affect the TTO and SG method differently, with loss aversion exerting an upward bias on both methods, but utility curvature only affecting TTO while probability weighting only affects SG.

Given the increased importance of CUA in informing health policy (Drummond et al., 2015), it is imperative to validly determine the weights that are ascribed to the relevant health states. The valuation of these health states, for example when obtaining tariffs for the commonly used EQ-5D-5L generic utility classification system (Versteegh et al., 2016), would necessarily occur within a descriptive context (Bleichrodt et al., 2001). This means that the status quo of applying EU and/or the linear QALY model to derive TTO and SG weights a

will not capture actual preferences, as these may include for example loss aversion, and b) may lead to different TTO and SG weights according to Bleichrodt (2002)²³. As such, our main motivation is to address the discrepancy between TTO and SG weights by obtaining these QALY weights using derivations based on a descriptively valid but non-normative theory (PT). We will refer to this process, where TTO and SG weights are obtained while incorporating loss aversion, non-linear utility and/or probability weighting into their derivation, as *correction for PT*. If correcting TTO and SG for PT is feasible, it could be used to correct observed responses in health state valuations, allowing corrected weights to be used when calculating QALYs to express health benefits in CUAs, as commonly done.

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

Our study features several methodological improvements compared to previous attempts at correcting TTO and/or SG weights for PT (see the Online Supplements of this dissertation for an overview). First, our adaptation of the non-parametric method (Abdellaoui et al., 2016) enables us to determine utility curvature, loss aversion and probability weighting separately for each individual, without assuming a specific parameter or parametrical form for these functions (as opposed to work by Martin et al., 2000, van der Pol and Roux, 2005, van Osch et al., 2004). We believe this is relevant, as large heterogeneity typically exists for PT elicitation (Pinto-Prades and Abellan-Perpiñan, 2012), warranting an individual measurement approach. Furthermore, applying specific parametric forms within experimental elicitation can confound results (Abdellaoui, 2000), thus allowing considerable bias to remain after correction (Wakker, 2008, Wakker, 2010). Second, we attempt to append the heterogeneity surrounding RPs by providing all subjects with the same RP, which is a hypothetical expected life duration (following the successful procedure described in Attema et al., 2013). This is important, since even though reference-dependence appears to be the most central tenet of PT, earlier work on the location of the RP suggests that individuals use multiple different health outcomes as RP (Bleichrodt et al., 2001, van Osch and Stiggelbout, 2008, van Osch et al., 2006, van Osch et al., 2004).

²³ These statements hold regardless if one believes EU to be the normative standard (as Kahneman & Tversky, 1979; Wakker, 2010 do), which would, for example, classify loss aversion as 'irrational' or a bias. We will make no such claims, and will refer to deviations of EU and the linear QALY model as generating bias in TTO and SG.

Theoretical framework

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

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

Several preliminaries are required before defining our full model (Eq. 7.1 and Eq. 7.2). We assume that preferences for health profiles are defined relative to a reference point (RP), which we denote as (β_r, T_r) . Following Wakker (2010), we define this RP as a point of comparison, that may differ during different parts of the analysis. Given that no plausible theory of RP selection is available (Wakker, 2010), we let the RP depend on framing of the decision context. Hence, (β_r, T_r) refers to an expected health profile described in a decision task, which is taken as the neutral point. This health profile has health status β_r , endured for T_r years. Throughout, for brevity, we denote the duration of all other health profiles as deviations from the RP, i.e. we denote health profiles (β_x, T_x) as (β_x, T_x^*) with $T_x^* = T_x - T_r$ in β_x . We will restrict our model to health profiles $(\beta_x, T_x^*) \succcurlyeq D$ with $\beta_x \succcurlyeq \beta_r$ for any T_x^* . In other words, we assume our model holds for a restricted outcome domain including only health profiles weakly preferred to immediate death, where health status remains at β_r or is improved.

Within this outcome domain, we model PT by incorporating sign-dependence for life duration, i.e. by modifying $L(T)$ in the general QALY model to $L^i(T^*)$. In our model, $L^i(T^*)$ is a standard, real-valued ratio scale utility function with $L^+(T_r) = 0$, which may be different

for gain outcomes $(\beta_x, T_x^*$, with $\beta_x \sim \beta_r$ and $T_x^* \geq 0$) and loss outcomes $(\beta_x, T_x^*$, with $\beta_x \sim \beta_r$ and $T_x^* < 0$). We do not modify $U(\beta)$ in our model, which implies that changes in health status will be evaluated as in the conventional general QALY model. We incorporate loss aversion²⁴ by taking $L^-(T^*) = \lambda L^+(T^*)$ for $T^* < 0$. Here λ denotes a loss aversion index, with $\lambda > 1$ [$\lambda = 1, \lambda < 1$] indicating loss aversion [loss neutrality, gain seeking]. Furthermore, we incorporate non-linear weighting of probabilities by incorporating probability weighting functions $w^i(p)$, $i = +, -$, for gains and losses respectively, that assign a number to each probability p , with $w^i(0) = 0$ and $w^i(1) = 1$.

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

$$w^+(p)U(\beta_x)L^+(T_x^*) + w^-(1-p)U(\beta_y)L^-(T_y^*), \quad (7.1)$$

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

$$w^i(p)U(\beta_x)L^i(T_x^*) + (1 - w^i(p))U(\beta_y)L^i(T_y^*), i = +, - \quad (7.2)$$

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

SG and TTO correction for PT

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

Our correction for PT involves deriving TTO and SG weights by means of our theoretical model based on PT. The application of our theoretical model requires assumptions about the RP used in TTO and SG. Typically, TTO and SG exercises are framed with the impaired health state (β_x, T_x) as RP. Furthermore, earlier work on SG²⁵ has suggested that the outcome

²⁴ In our simplified approach, we model PT over life duration by assuming attribute-specific evaluation (as in Bleichrodt et al., 2009). Loss aversion is, thus, defined over life duration, as it is not meaningful on $U(\beta_x)$ when health status is considered a qualitative measure (Bleichrodt and Miyamoto, 2003). This does not affect our analysis, since we only consider improvements in health status.

²⁵ No empirical work exists studying the RP for TTO. Here we assumed that it coincides with that of SG, and with how TTO is typically framed. If the time spent in perfect health (i.e. FH, T_y) is taken as RP instead, Eq. 7.3 cannot be applied. This also holds for SG, i.e. Eq. 7.4 is only valid if the RP is actually (β_x, T_x)

that remains constant, i.e. the time spent with reduced health status (β_x, T_x) usually is taken as RP (Bleichrodt et al., 2001, van Osch et al., 2006). Hence throughout the paper we will make the following assumption about the RP for TTO and SG: $(\beta_r, T_r) = (\beta_x, T_x)$.

Under these assumptions, TTO indifference $(\beta_x, T_x) \sim (FH, T_y)$ allow the following derivation for $U(\beta_x)$ ²⁶:

$$U(\beta_x) = \frac{L^-(T_y^*)+1}{(1-\lambda)L^-(T_y^*)+1}, \quad (7.3)$$

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

$$U(\beta_x) = \frac{w^+(p)}{w^+(p) + \lambda w^-(1-p)}. \quad (7.4)$$

Parameter elicitation

In order to correct both TTO and SG weights for PT, i.e. to be able to compute the outcome of Eq. 7.3 and Eq. 7.4, one needs to elicit: a) $L^i(T^*)$ with T_x^* as RP to allow estimation of $L^-(T_y^*)$, b) probability weighting functions $w^i(p)$, $i = +, -$, and c) a loss aversion coefficient λ which reflects overweighting of losses with T_x^* as RP. This means that t_x should be kept constant across TTO and SG and the elicitation of $L^i(T^*)$, to ensure that λ refers to the same theoretical construct throughout (i.e. the same kink around the RP).

Methods

We report the results of an experiment in which we compare TTO and SG weights derived assuming EU and the linear QALY model to QALY weights corrected for PT (i.e. by Eq. 7.3 and Eq. 7.4). In this experiment PT parameters were elicited using methodology based work by Abdellaoui and colleagues (2016). To reduce the influence of order effects and test for consistency, multiple counterbalancing procedures were conducted between participants and consistency checks were in place (see the Online Supplements of this dissertation). The experiment was computerized in Matlab. Subjects were 99 students of the Rotterdam School of Management (58 female), who were rewarded course credits. Experimental sessions lasted for approximately 55 minutes and were run on computers in sessions of four subjects sitting adjacently in separate cubicles. An instructor was present at all times to answer questions.

TTO and SG weight elicitation

We elicited TTO and SG weights for a total of four health states (one practice state) from the EQ-5D-5L descriptive system (Herdman et al., 2011). These health states reflected an array of mildly aversive health states, in order to avoid health states that could be considered worse than death (Dolan, 1997). The following health states were used: 22222 (practice, β_p), $\beta_1 = 21211$, $\beta_2 = 31221$, and $\beta_3 = 32341$. We applied a bisection choice-based elicitation

²⁶ Eq. 7.3 and Eq. 7.4 apply a scaling of $L^i(T^*)$, where the utility of the lowest outcome is set to -1, for simplicity (i.e. $L^-(T_a) = -1$). For elaborate proofs of Eq. 7.3 and Eq. 7.4 under our theoretical model see the Online Supplements of this dissertation.

procedure with four consecutive choices, as choice-based procedures produce more consistent measurements than matching (Noussair et al., 2004). Subjects were asked to imagine having lived until age 50 in perfect health after which they contracted a disease which would affect their quality of life for their remaining life expectancy of 20 years. TTO and SG were completed for these remaining 20 years (i.e. $t_a = 50$). In both cases the maximum expected age of death was 70 years, i.e. subjects made decisions with regard to the quality of life for age 50 to 70 (followed by death), which ensured that t_x was constant for both TTO and SG.

Non-parametric method

We adapted Abdellaoui and colleagues' (2016) non-parametric methodology to measure PT under risk in the health domain. In order to elicit $L^i(T^*)$ with the same t_x as RP as in TTO and SG, we instructed subjects to take living from current age until 70 in perfect health as RP, i.e. $(\beta_r, T_r) = (FH, 70 - t_a)$. Elicitation consisted of four stages (an elaborate description of the method and instructions can be found in the Online Supplements of this dissertation). The first stage connected utility for gains ($L^+(T^*)$) to the utility for losses ($L^-(T^*)$). The second and the third stages employed the trade-off method of Wakker and Deneffe (1996) to measure a standard sequence of utility for gains and utility for losses, respectively. The fourth stage measured probability weighting, separately for gains and losses, i.e. $w^+(p)$ and $w^-(p)$. Our methodology thus makes it possible to completely elucidate PT's tenets in the health domain, without imposing parametric assumptions on $L^i(T^*)$ and $w^i(p)$. Each of the four stages had slightly different instructions (see the Online Supplements of this dissertation), providing the context for the trade-offs subjects were required to make. Subjects had to choose between two medicines which could amend their situation, but would not affect their life expectancy, which remained constant at perfect health. All indifferences were elicited using a bisection choice-based procedure with a slider (following Abdellaoui et al., 2016) where subjects first performed three binary choices. This procedure zoomed in to the point at which subjects would become indifferent, but still allowed subjects to specify the final value and adjust accordingly. To allow estimation of $L^-(T_y^*)$ in Eq. 7.3 regardless of the amount of years given up in TTO, subjects' standard sequence continued to at least 20 years above and below t_x (i.e. living until 70), to avoid extrapolation beyond the measured curve²⁷.

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

We used two methods to investigate the curvature of $L^i(T^*)$, i.e. utility curvature: a non-parametric and a parametric method (similar to Abdellaoui et al., 2016). For these analyses of utility curvature, we normalized all durations by dividing through subjects' highest absolute elicited duration for gains and losses, respectively (T_{kG}^* or $-T_{kL}^*$). This resulted in T^* being in the range $[-1, 1]$. Next, we calculated the area under the curve (AUC) of $L^i(T^*)$ separately for both domains, by setting $L^+(T_{kG}^*) = 1$ and $L^-(T_{kL}^*) = -1$. If utility of life duration is linear,

²⁷ After 25 steps the standard sequence elicitation was terminated, to avoid overburdening our subjects. When necessary, $L^-(T_y^*)$ was obtained by extrapolation.

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 $> 1/2$, concave $< 1/2$). This method of analyzing utility curvature is non-parametric. We also analyzed $L^i(T^*)$ parametrically by employing the most commonly used power utility family using non-linear least squares, using the same normalizations. For this family, $L^+(T^*) = (T^*)^\alpha$ and $L^-(T^*) = -(-T^*)^\alpha$ with $\alpha > 0$. For gains [losses], $\alpha > 1$ corresponds to convex [concave] utility, $\alpha = 1$ corresponds to linear utility, and $\alpha < 1$ corresponds to concave [convex] utility.

Analyses of loss aversion

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

Probability weighting

We used certainty equivalences using varying probabilities to elicit the weighting functions, similar to Attema and colleagues (2018a). In particular, we used linear interpolation to obtain a $w^+(p)$ and $w^-(p)$, using $p = 0.1, 0.3, 0.5, 0.7, 0.9$. Furthermore, we used Tversky and Kahneman's (1992) one-parameter inverse S-shaped probability weighting function $w^i(p) = p^\gamma / (p^\gamma + (1-p)^\gamma)^{1/\gamma}$ with $i = +, -$, estimated by nonlinear least squares. The γ -parameter controls for the shape of the probability weighting function. If $\gamma = 1$ there is no probability transformation and $w^i(p) = p$. However, if $\gamma < 1$, decision makers underweight large probabilities and overweight small probabilities. This corresponds to the commonly found inverse S-shaped weighting function. If $\gamma > 1$, the opposite pattern holds, corresponding to an S-shaped weighting function.

Results

Two subjects expressed unwillingness to trade off any life years, which caused the experiment to fail. These subjects were removed from further analyses. As can be seen in the Online Supplements of this dissertation, 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). 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). Correlations between consistency checks and original values were high, suggesting strong association between these values (Kendall's τ 's > 0.51 , p 's < 0.003)

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

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

We observed 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 function to the data, we found a median α of 0.787 for gains and 0.757 for losses (significantly smaller than 1, Wilcoxon signed ranks test: $p < 0.001$). Thus, both parametric and non-parametric results demonstrated $L^+(T^*)$ to be concave and $L^-(T^*)$ to be convex.

Table 7.1 shows the classification of subjects' curvature for gains ($L^+(T^*)$) and losses ($L^-(T^*)$) at the individual level, both parametrically and non-parametrically. The most common pattern was concave curvature for $L^+(T^*)$ and convex curvature for $L^-(T^*)$ as was found in an earlier implementation of this method (Attema et al., 2018a). This conclusion holds for both non-parametric (53%) and parametric (53%) results.

Table 7.1. Classification for curvature of $L^+(T^*)$ and $L^-(T^*)$ at the individual level

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

Loss aversion

Utilizing Köbberling and Wakker’s (2005) definition, we found a median loss aversion index of $\lambda = 2$ (IQR:1.00 – 3.52). Thus, we found considerable loss aversion at the aggregate level, with the median being significantly higher than 1 (Wilcoxon test: $p < 0.001$). At the individual level, the majority of subjects demonstrated loss aversion, with 72% ($n = 70$) classifying as loss averse, 15% ($n = 15$) and 13% ($n = 12$) classifying as loss neutral or gains seeking, respectively.

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

Figure 7.1 shows the median decision weights assigned to $p = 0.1, 0.3, 0.5, 0.7, 0.9$. As can be seen from the plots, we observe inverse S-shaped probability weighting both for gains and losses, with more pronounced overweighting of small probabilities for losses. Using Tversky and Kahneman’s (1992) one-parameter function, we found a median $\gamma = 0.92$ for gains and a median $\gamma = 0.84$ for losses (both significantly lower than 1, Wilcoxon test: p ’s < 0.04). Both analyses demonstrated that the typical inverse S-shaped probability transformation was the most prevalent in 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.

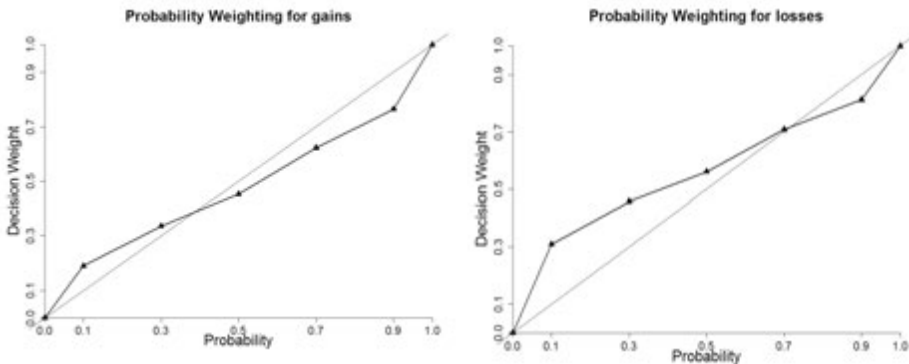


Figure 7.1. Probability weighting functions for gains ($w^+(p)$) and losses ($w^-(p)$)

Health state correction

Table 7.2 shows QALY weights for all health states elicited using TTO and SG, where uncorrected refers to weights elicited assuming EU and linear QALYs, while corrected weights are elicited by means of Eq. 7.3 and Eq. 7.4. To test the sensitivity of our results to linear interpolation, we also corrected TTO and SG weights by using power utility to estimate $L^-(T_y^*)$, and the Tversky and Kahneman’s (1992) probability weighting function to estimate $w^+(p)$ and $w^-(1 - p)$, these are indicated by ‘Parametric Corrections’ in Table 7.2. 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 β_x . Our results show that the corrected weights were

lower than the uncorrected weights for TTO and SG (paired t-test: all p 's < 0.01). The initially significant difference between the uncorrected weights only disappeared for all β after applying non-parametric corrections (paired t-test: all p 's $> .09$). The parametric corrections left significant and substantial differences between TTO and SG weights.

Finally, we performed four isolated corrections. For the sake of brevity, we only report the results of the non-parametric corrections (see the Online Supplements of this dissertation for results of these analyses for parametric corrections). First, we corrected TTO for utility curvature only, with $\lambda = 1$. Second, TTO weights were corrected for loss aversion only, with linear utility (i.e. $L^i(T^*) = T^*$). Third, we corrected SG for probability weighting only, with $\lambda = 1$. Finally, SG weights were corrected for loss aversion only, with $w^i(p) = p$. This allowed us to demonstrate the influence of each correction in isolation. Table 7.3 shows that correcting for loss aversion had a stronger downward influence on TTO weights than correcting for curvature of $L^i(T^*)$, and both correcting for probability weighting and correcting for loss aversion had a substantial negative influence on SG weights.

Table 7.2. Overview of mean weights [standard deviation] for health states β_{1-3} for TTO and SG including differences between methodologies under multiple corrections

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

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

Table 7.3. Isolated effects of corrections for utility curvature (UC), loss aversion (LA) and probability weighting (PW) for TTO and SG weights [standard deviation in brackets]

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

Discussion

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

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

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

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

Several limitations of our study need noting. First, several subjects violated monotonicity for the health states used. Although excluding these subjects from the sample did not alter our results, we expect that these errors in decision-making are to be attributed to either a) imprecision of preferences or b) error propagation, i.e. early errors cascading into later stages of the task. Considering the use of only relatively mild health states, for which subjects may have no precise preference ordering in mind, some overlap may occur within our method. Regarding error propagation, it is good to note that during utility elicitation, subjects could rectify errors by adjusting the final indifference value on the slider to any non-dominant value in life years, i.e. fix their earlier ‘errors’. Testing for error propagation, by performing an error simulation as described by Bleichrodt and Pinto (2000), confirmed that errors did not have a propagating effect on the standard sequence we elicited for gains and losses²⁸.

Second, concerns may be raised about the role of the RP in this paper. We find that the observed discrepancies between TTO and SG can be removed by correcting under the assumption that decision makers utilize the guaranteed outcome (β_x, T_x) as RP (which ensures that t_x remains constant). However, earlier work on health-related preferences has suggested that individuals may also use their own current health and life expectancy as RP (van Nooten and Brouwer, 2004, van Nooten et al., 2009). In our work, we found no evidence of such effects²⁹. A related limitation concerns our assumption that subjects use the fixed outcome in both TTO and SG as their RP which is crucial for our results as our corrections depend on a constant T_r throughout the multiple parts of the experiment. Earlier work, however, demonstrated that SG subjects may also use the time spent in full health as their RP (van Osch and Stiggelbout, 2008). To our knowledge, such work does not exist for TTO

²⁸ The difference between TTO and SG weights was not significant in all simulations (k=1000) for β_1 and β_2 , whilst replicating our results in the majority of simulations for β_3 (over 70%). These simulations suggest that our correction method is quite robust to error propagation.

²⁹ We tested for associations between subjects' self-reported life expectancy and their estimates for loss aversion, utility curvature and probability weighting, nor were such associations observed for raw and corrected health state weights (all Kendall's τ 's < 1.52, all p's > 0.13).

methods. Therefore, future work should explore the possibility of correcting under the assumption that subjects use full health as RP, both for TTO and SG.

Finally and perhaps most importantly, the primary goal of the present research was merely to provide the first empirical test of Bleichrodt's (2002) predictions for TTO and SG weights, and our findings should be interpreted in this context. We observed considerable differences to nationally representative findings. For example, the Dutch tariff (Versteegh et al., 2016) for health state β_1 (21211) is 0.876, while we elicited a raw TTO weight of 0.665. Our sample, consisting of young, healthy students will have contributed strongly to this initial discrepancy, next to differences in methodology. We also note that after correction, the discrepancy between tariffs and corrected weights increases. After the non-parametric correction, the QALY value of state β_1 decreases to 0.492. Clearly, this calls for further investigation of the methods used here, also in other (general public) samples, in order to further explore the impact of corrections and further refine the methods used. This future research may also clarify whether our framing may have yielded relatively low weights and how the methods used here can be simplified to be suitable for use in general public samples.

Conclusion

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

8

The corrective approach:
policy implications of
recent developments in
QALY measurement based
on prospect theory.

Chapter based on:

Lipman, S. A., Brouwer, W. B. F., & Attema, A. E. (2019). The corrective approach: policy implications of recent developments in QALY measurement based on prospect theory. Value in Health, 22 (7), 816-821.

Abstract: Common health state valuation methodology, such as time trade-off (TTO) and standard gamble (SG), is typically applied under several descriptively invalid assumptions, for example related to linear Quality-Adjusted Life Years (QALYs) or expected utility (EU) theory. Hence, the current use of results from health state valuation exercises may lead to biased QALY weights, which may in turn affect decisions based on economic evaluations using such weights. Methods have been proposed to correct responses for the biases associated with different health state valuation techniques. In this paper we outline the relevance of prospect theory (PT), which has become the dominant descriptive alternative to EU, for health state valuations and economic evaluations. We provide an overview of work in this field, which aims to remove biases from QALY weights. We label this ‘the corrective approach’. By quantifying PT parameters, such as loss aversion, probability weighting and non-linear utility, it may be possible to correct TTO and SG responses for biases, in an attempt to produce more valid estimates of preferences for health states. Through straightforward examples this paper illustrates the effects of this corrective approach, and discusses several unresolved issues that currently limit the relevance of corrected weights for policy. Suggestions for research addressing these issues are provided. Nonetheless, if validly corrected health state valuations become available, we argue in favor of using these in economic evaluations.

Introduction

Health economic evaluations provide important information to policy makers (Drummond et al., 2015), for example by determining incremental cost-effectiveness ratios (ICERs) of interventions, i.e. the incremental costs per unit of health gained. In cost-utility analyses health gains are commonly expressed in Quality-Adjusted Life-Years (QALYs), which are obtained by multiplying life duration with the utility weight(s) of the health state(s) experienced. These QALY weights are normalized such that 0 and 1 represent the utility of health states judged equivalent to being dead or perfect health, respectively. It is well-known that QALY weights differ between health state valuation (HSV) methods used to obtain them: standard gamble (SG) weights are typically higher than weights obtained with time trade-off (TTO) methodologies (e.g. Bleichrodt and Johannesson, 1997, Read et al., 1984, Torrance, 1976). Bleichrodt (2002) proposed that these differences occur as result of bias due to the ‘classical elicitation assumption’, i.e. applying expected utility (EU) theory to analyze individual choices (2001). Although research in behavioral economics and psychology has established many systematic violations of EU (for a review of these violations in the monetary domain, see Starmer, 2000), its axioms still underlie QALY weight calculations applied in HSV exercises (Bleichrodt et al., 2007, Llewellyn-Thomas et al., 1982, Treadwell and Lenert, 1999). In order to better inform healthcare decisions, it has been suggested that these biased QALY weights could be corrected, by applying calculations based on alternative utility models such as prospect theory (Bleichrodt et al., 2001, and Chapter 7 of this dissertation).

Prospect theory (PT) (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992) by now is a well-established behavioral theory, which assumes that people judge states relative to some reference point (such as the current position). Changes relative to that point are perceived as either losses or gains. Furthermore, utility increases for gains are lower than utility decreases for equally sized losses, i.e. people are loss averse. People, moreover, are not ‘perfect calculators’. They tend to overweight small probabilities and underweight large ones. This is labeled probability weighting (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992). It has been suggested that reference points, loss aversion and probability weighting affect decisions about health (e.g. Attema et al., 2016, Bleichrodt et al., 2007, Bleichrodt and Pinto, 2005, Jonker et al., 2017, and Chapter 7 of this dissertation), perhaps more pronouncedly than in financial decision-making (Suter et al., 2016). Importantly, these insights may provide an explanation for the systematic differences between HSV methods (Bleichrodt, 2002), and can be used in pursuit of obtaining QALY weights that may more accurately reflect trade-offs relevant to specific methodologies. For instance, SG responses can be corrected for bias due to probability weighting, and TTO responses can be corrected for loss aversion. It has been argued that such a ‘corrective approach’ may lead to better (i.e. less biased) QALY weights and hence may have relevance for HSV (Bleichrodt, 2002, and Chapter 7 of this dissertation). TTO and SG (see Box 8.1 for examples) have been the focus of the corrective approach, as these methods are especially relevant to HSV for generic utility classifications, with EQ-5D tariffs frequently being determined via TTO (e.g. Versteegh et al., 2016), while SF-6D tariffs have been obtained via SG (Walters and Brazier, 2005).

Hence, in this paper, we focus on the corrective approach in the context of TTO and SG, by providing an overview of developments in the corrective approach. These developments opened up at least two challenges to research and policy, which are discussed in the next section. First, applying the corrective approach (with current estimates) may affect ICER calculations and allocation decisions – especially when perfect health is involved. Second, even though loss aversion may lead to bias in HSV, it could reflect a real preference many individuals may hold. Thus, distinguishing between gains and losses may still be seen as relevant in health care decision making. Preventing health losses may, for example, have higher societal value than achieving health gains of a similar size (relative to a relevant reference point). Hence, we explore how a loss aversion premium for prevented health losses could be applied if and when deemed relevant by responsible policy makers. Finally, we outline policy implications and important steps for future research.

Box 8.1 The Time Trade-Off (TTO) and Standard Gamble (SG) methods

TTO exercises involve choices between living longer (say 10 years) in a poor health state or shorter ($x < 10$) in perfect health. Assuming the linear QALY model, the utility of the imperfect health state is given by $x/10$. The number of years in X is varied until the respondent is indifferent between the two options. Hence, if a person considers 6 years in perfect health to be equal to 10 years with severe pain, the utility of this health state is $6/10 = 0.6$. The worse the health state, the greater is the reduction in years in perfect health that people would be willing to accept. Similarly, SG methods entail asking subjects to choose between living some period of time (e.g. 10 years) in some imperfect health state for sure and a gamble with two outcomes: full health (FH) for the same period of time, or immediate death (D). By varying the probability of immediate death, one may derive the utility of the imperfect health state. Under EU (and with the utility of perfect health normalized to 1 and that of death to 0) this utility equals probability $1-p$. For instance, if people accept a maximum risk of 10% of immediate death to live the rest of their lifespan in perfect health rather than with moderate back pain, this implies the utility of the health state ‘moderate back pain’ is 0.9. If the health state is worse, people would accept a higher risk of immediate death to regain health, leading to lower QALY weights.

The corrective approach: rationale and overview of earlier work

Acknowledging that decisions about health may be reference-dependent, as is done in prospect theory, changes the implications of responses in TTO and SG exercises (Bleichrodt, 2002). These implications crucially depend on the location of the reference point in HSV exercises, which was the topic of some empirical studies (e.g. van Osch et al., 2006). Given that this work suggested that the time spent in the imperfect health state was the most frequently applied reference point (coinciding with how TTO and SG are typically framed), we will assume this reference point in HSV throughout this paper. Under this assumption, TTO involves trading off losses in life duration for gains in quality of life, whereas picking the risky option in SG indicates a preference for a mixed gamble generating either a gain in quality of life or a catastrophic loss of life (i.e. immediate death). As such, loss aversion may exert upward bias in both methods, because the negative utility of (possible) losses subjects are willing to incur in TTO and SG is amplified by loss aversion and, thus, signifies a larger utility decrement than assumed under the classical elicitation assumption (Bleichrodt, 2002).

Probability weighting only affects SG, and generally has an upward influence on SG weights. This upward bias results from subjects' overweighting of generally small chances of death, and underweighting of the typically large chance of obtaining full health in this method (Bleichrodt, 2002). In other words, if subjects weight probabilities in this manner, accepting a 10% chance of death in SG may signify a larger utility decrement than traditionally assumed. Additionally, linear utility of life duration is often assumed in health state valuation (i.e. the linear QALY model). However, many authors have found utility of life years to deviate from linearity in the ranges typically considered in TTO (e.g. Abellan-Perpignan et al., 2006, Attema et al., 2012, Bleichrodt and Pinto, 2005, Wakker and Deneffe, 1996), where the severity of this deviation may even depend on how duration is described (Craig et al., 2018). Such utility curvature will only affect TTO weights, as this method depends critically on trade-offs in duration. As shown in (Bleichrodt, 2002), if utility of life years is concave (i.e. each extra year of life is worth less) instead of linear, TTO weights are biased downwards. Inversely, when utility of life years is convex (i.e. each subsequent year is worth more than the previous) instead of linear, then the TTO weights are biased upwards.

Although EU is often considered the 'right' normative theory (Harsanyi, 1955, Kahneman and Tversky, 1979, Savage, 1954, Wakker, 2010), retaining the classical elicitation assumption mistakes the empirical nature of HSV, in which deviations from EU are likely if not inevitable (Bleichrodt et al., 2001), with the normative relevance that QALYs may have in economic evaluations. Consequently, several studies exist that applied a corrective approach (Attema and Brouwer, 2009, Martin et al., 2000, Oliver, 2003a, Perpiñan et al., 2009, Stiggelbout et al., 1994, van Osch et al., 2004, Wakker and Stiggelbout, 1995), each using the same two steps: 1) quantify the deviations from EU and the linear QALY model, such as loss aversion and non-linear utility, and 2) utilize corrective formulas (e.g. Attema and Brouwer, 2009, Bleichrodt et al., 2001, and Chapter 7 of this dissertation) to account for their confounding effect on HSV. Considerable differences exist between empirical studies regarding both steps, with researchers using different techniques to quantify PT, and/or applying corrective formulas based on different assumptions about decision-making. A frequently applied approach is to preemptively assume a certain degree of loss aversion, utility curvature and probability weighting in all respondents (e.g. Pinto-Prades and Abellan-Perpiñan, 2012). In this type of work, average parameters elicited in earlier work (e.g. loss aversion coefficients of 2.25) are applied to each individual. However, typically large differences in loss aversion, utility curvature and probability weighting are observed between individuals, i.e. not everyone is equally loss averse or weighs probabilities the same way.

Therefore, other attempts at correcting TTO and/or SG weights apply an individual approach, in which PT parameters are elicited separately for each respondent, applying corrections for loss aversion (e.g. Oliver, 2003a) or non-linear utility of life duration (e.g. Attema and Brouwer, 2009, Stiggelbout et al., 1994), for example. In this work, utility of life duration or probability weighting are typically estimated by assuming specific functional forms (Martin et al., 2000, Stiggelbout et al., 1994, van der Pol and Roux, 2005, van Osch et al., 2004). Although such parametric analyses may be practical and efficient, the mathematical properties of the chosen parametrical form may not fit well for some extreme cases (for an example, see Wakker, 2008). Indeed, a literature exists documenting that parametric analysis may result in biases in individual estimates for PT (Abdellaoui et al., 2016, Abdellaoui et al., 2007). In Chapter 7 of this dissertation, Abdellaoui and colleagues' (2016) non-parametric method was adapted to correct TTO and SG weights without parametric assumptions. In that

study, as was expected under PT, concave utility for life year gains and convex utility for losses was observed, with considerable loss aversion and probability weighting for both gains and losses. After applying the corrective approach, TTO and SG weights converged, as predicted by Bleichrodt (2002). However, the resulting corrected QALY weights seemed quite low and compressed, raising questions about their validity (see the Online Supplements of this dissertation for numerical examples of corrections based on this study).

Collectively, these developments in PT measurement and the corrective approach could be important for health policy, as they suggest that it may be possible to move beyond the classical elicitation assumption for TTO and SG weights, which still dominates applications of HSV.

The impact of the corrective approach on health policy

Regardless of these developments, the corrective approach currently does not affect the policy domain: only a single study (Perpiñán et al., 2009) exists that estimated corrected tariff lists (i.e. without assuming EU), and no country has adopted the corrective approach in guidelines for economic evaluations. Of course, this gap between the current state-of-the-art in the literature and policy may in part be caused by unresolved questions about validity or feasibility of the corrective approach. We return to these important questions in a subsequent section, for now disregarding them in order to address two currently understudied corollaries of applying the corrective approach. First, we illustrate with currently available weights that moving from the classical elicitation assumption to a corrective approach may substantially affect ICERs and allocation decisions, especially when treatments involve perfect health. Second, we explore how loss aversion, which produces bias that we argued needs correction in HSV, could still have relevance in the context of health policy.

Box 8.2. The impact of the corrective approach on ICERs

Imagine a group of patients who experience moderate problems with walking about, slight problems with usual activities and slight pain or discomfort (31221 in EQ-5D nomenclature, β_2 in the Online Supplements of this dissertation). In Chapter 7 of this dissertation, the classical TTO and SG weights for β_2 were elicited at 0.605, and 0.706, respectively. We let $U(\cdot)$ represent the utility assigned to health states. Assume that a treatment is evaluated that returns these patients to full health for 30 years, and the costs for treatment are € 20,000 per year. Without discounting, we then obtain the following ICERs:

$$ICER_{TTO} = \frac{\text{€}20,000 * 30 \text{ years} = \text{€}600,000}{30 * (U(FH) - U(\beta_2)) = 30 * 0.395 = 11.85} = 50,632\text{€/QALY},$$

$$ICER_{SG} = \frac{\text{€}20,000 * 30 \text{ years} = \text{€}600,000}{30 * (U(FH) - U(\beta_2)) = 30 * 0.294 = 8.82} = 68,027\text{€/QALY}.$$

If we repeat our calculations using corrected SG and TTO weights, which were 0.442 and 0.456 respectively (see the Online Supplements of this dissertation), we obtain the following ICERs:

$$ICER_{TTO-c} = \frac{\text{€}20,000 * 30 \text{ years} = \text{€}600,000}{30 * (U(FH) - U(\beta_2)) = 30 * 0.558 = 16.74} = 35,842\text{€/QALY},$$

To correct or not to correct: it makes a difference!

Currently, TTO and SG weights (or weights derived from classification systems using these methods) are commonly elicited assuming EU and/or the linear QALY model. Hence, at least implicitly, the classical elicitation assumption is still applied. Our focus is to compare this status quo to the situation in which the corrective approach would be applied. We will refer to TTO and SG weights calculated under the classical elicitation assumption as classical weights, and refer to corrected weights when the corrective approach is applied. Without correction, TTO and SG typically yield different QALY weights (e.g. Bleichrodt and Johannesson, 1997, Read et al., 1984, Torrance, 1976), and hence, it is obvious that ICERs for the same treatment could vary substantially (and systematically) depending on which method is utilized to value health benefits – especially for treatments dealing with full health. If we choose to apply corrections, we could observe converging TTO and SG weights, and hence converging ICERs for both methods (see Box 8.2 for an example using currently available estimates). Similarly, applying the corrective approach may affect allocation decisions in different situations compared to using classical weights (see Box 8.3 for an example). In both cases, applying the corrective approach will likely lead to a lower valuations of impaired health states (see Chapter 7 of this dissertation). Although applying a corrective approach in calculations of QALY weights from HSV exercises will likely improve understanding of choices in TTO and SG, it is not yet clear to what extent this corrective approach ultimately yields QALY weights that better reflect preferences for health states. Obviously, the exact impact of utilizing corrected weights instead of classical weights on subsequent economic evaluations will depend on the respective valuations of health states associated with the treatment and control groups, which raises two crucial issues. First, whereas Box 8.2 and Box 8.3 illustrate that the corrective approach may have considerable effects on ICERs and allocation decisions for treatments moving patients between impaired health states and full health, the effects of correction on movements between health states that differ only slightly is currently unknown. Given that treatments yielding full recovery are likely to be rare, more insight into how such small improvements or deteriorations in health status are affected by deciding to correct or not to correct (for both SG and TTO) is an important avenue for future research. Second, as can be seen from Box 8.2 and Box 8.3, another corollary of applying the corrective approach is that a ‘perfect-health gap’ may be exacerbated. Note that whether or not such a gap emerges depends on final corrected weights. However, currently available corrected weights suggest that the distance in utility between the mildest impaired health states, which in current estimates receive lower QALY weights after correcting for bias due to loss aversion, utility curvature and probability weighting, and the utility of perfect health, which remains stable at 1.00, increases. As a result, applying the corrective approach may especially impact ICERs of and allocation decisions for treatments involving patients losing or returning to perfect health. Incremental cost-effectiveness thus increases (as shown in Box 8.2) for treatments that return patients to full health, with potential policy and allocation implications (as in Box 8.3).

It is yet unclear whether this ‘perfect health gap’ is simply the result of poor correction of bias in TTO and SG, and as such an unintended and undesirable by-product of applying the corrective approach, or reflects actual individual or societal preferences. Implicitly, a perfect health gap already exists in many applications of tariff estimations for utility classification systems (e.g. Kim et al., 2016, Versteegh et al., 2016, Xie et al., 2016). Whether the larger

gap aligns with preferences needs to be established further, especially since correction may enlarge the gap, emphasizing the special status perfect health may have. However, earlier work applying a similar corrective approach outside the health domain found that correcting for PT may lead to compression of utility weights (Bleichrodt et al., 2001). It was suggested that this compression was unrelated to individuals' preferences, but rather resulted from the specific parametrized correction process applied. Such compression of corrected weights could explain the enlarged perfect health gap. Indeed, if the utility estimates of impaired health states are compressed, and come closer to the midpoint of the 0 to 1 scale (as can be seen in the Online Supplements of this dissertation), while perfect health remains fixed at 1, this inevitably leads to a (larger) gap.

Moreover, if this compression effect is strong enough, it could also explain the convergence of TTO and SG valuations (as all values cluster in the middle of the scale). The convergence of valuations using both methods has been interpreted in earlier work as evidence of

Box 8.3. The impact of the corrective approach on allocation decisions

Imagine two patient populations that have the same initial quality of life: a mild health state characterized by slight problems in mobility and self-care (21211, β_1 in the Online Supplements of this dissertation). Treatment A will return population P_a to perfect health, i.e. Treatment A is curative. Treatment B, on the other hand, will prevent population P_b from a sure loss in quality of life, from state β_1 to β_3 (32341), a health state characterized by moderate problems with mobility, slight problems with self-care, moderate problems with usual activity, and severe pain. In other words, we have to choose between funding A, which involves P_a gaining $U(FH) - U(\beta_1)$, while funding B prevents a loss in quality of life of $U(\beta_3) - U(\beta_1)$ for P_b . Under the classical elicitation assumption, we observe that the utility differences between $U(\beta_3) - U(\beta_1)$ and $U(FH) - U(\beta_1)$ are of similar magnitude, independent of which method is used to elicit these weights (see Figure B8.3a). After correction, however, utility for β_1 (see Figure B8.3b) has dropped substantially, which could change the allocation decision problem between A and B in favor of Treatment A (ceteris paribus).

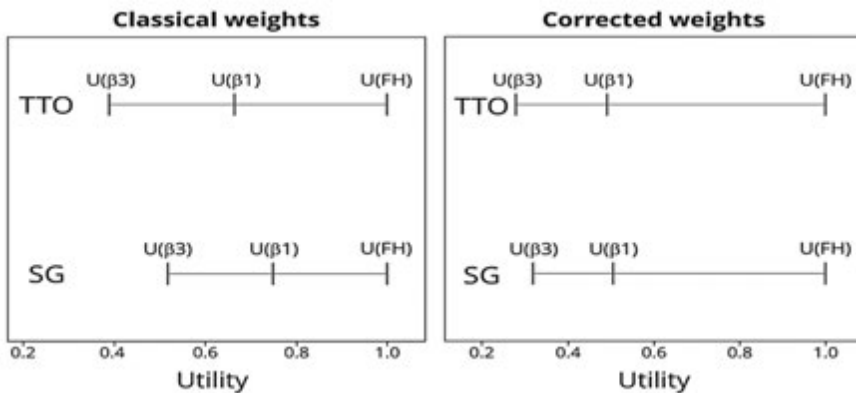


Figure B8.3a. TTO and SG weight differences with classical estimation

Figure B8.3b. TTO and SG weight under the corrective approach

successful correction (van Osch et al., 2004, and in Chapter 7 of this dissertation). Hence, it is crucial to determine whether the corrective approach leads to such unwarranted compression of QALY weights, and indeed whether corrected weights better reflect preferences for health states than classical weights. The move towards individual corrective approaches (see Chapter 7 of this dissertation), combined with for example ex-post validation of corrected weights in personal interviews in future work, could shed light on this issue. Such insight in the validity of classical and corrected weights is pivotal in interpreting the observed convergence of health state valuations obtained through different methods as well as the increased perfect health gap, and we believe is required before the corrective approach is applied in economic evaluation.

To prevent is better than to cure: exploring the loss aversion premium

Applying the corrective approach implies correcting for bias in TTO and SG weights that results from loss aversion, probability weighting and utility curvature. This may be desirable, because TTO and SG are not designed to reflect these time and risk preferences; they were designed to reflect preferences for health states. As such, in our view, if time and risk preferences are deemed relevant for health policy, HSV is not the context in which they should be considered. Rather, this should occur within economic evaluations if deemed appropriate. For time preferences this is already common practice: often a discount rate is applied to future life years in cost-effectiveness analyses (Drummond et al., 2015), which may reflect societal time preferences for health outcomes (Attema et al., 2018b). However, to avoid ‘double discounting’, TTO weights should be adjusted for individual utility curvature (or time preferences) before applying such societal discount rates in economic evaluations (MacKeigan et al., 2003). Thus, i) individuals’ discounting in TTO should be corrected for initially in HSV and ii) policy makers can decide if and which societal time preference is to be incorporated in economic evaluations. Application of the corrective approach would extend the first step of this sequence to also correct for loss aversion and probability weighting. However, no work exists on providing a rationale or methods for also applying the second step for loss aversion or probability weighting, even though it is well-known that loss aversion and probability weighting apply to health outcomes as well (Attema et al., 2018a, Attema et al., 2013, Attema et al., 2016, Kemel and Paraschiv, 2018, and Chapter 7 of this dissertation).

Several authors provided arguments that loss aversion and probability weighting, although yielding bias in TTO and SG, need not be irrelevant or irrational. For example, Huber and colleagues (Huber et al., 2002) wrote: ‘In many settings, one cannot tell whether loss aversion is a bias or merely a reflection of the fact that losses have more emotional impact than gains of equal magnitude’. Similarly, Diecidue and Wakker (Diecidue and Wakker, 2001) argued that probability weighting could reflect individuals’ decision that some outcomes are especially important and should receive more attention than equally likely outcomes. As such, just as societal time preference, both probability weighting and loss aversion *could* provide information, relevant for economic evaluations and health care decision-makers. They may signal that (possible) health losses are perceived to have large emotional impact by many members of society. Hence, this preference information could be viewed as a relevant input in decisions based on economic evaluations of health technologies dealing with (risks of) health losses. Below, we will explore how policy makers may include such behavioral insights in such economic evaluations, with a focus on loss aversion

(developing a similar approach for probability weighting is beyond the scope of this paper and less intuitive in the context of economic evaluation).

To interpret or apply insights based on loss aversion in economic evaluations, it is important to consider which reference point is taken – otherwise losses and gains are undefined. For example, one could take individuals' current health as reference point, which implies that preventive treatments reduce health losses, while curative treatments generate health gains (typically after some loss is incurred). Loss aversion could then refer to a social preference for preventive treatments over curative treatments (*ceteris paribus*). However, an extensive literature on equity weighting in health exists suggesting that people on average prefer to treat those worse off (e.g. Van de Wetering et al., 2013). Furthermore, research has also documented that age-dependent expectations about length and quality of life could also serve as reference point (Brouwer et al., 2005, van Nooten and Brouwer, 2004, Wouters et al., 2015). Collectively, these findings indicate that if a similar approach is to be developed as for time preference, more research on reference points in decisions about health is required. Nonetheless, in the Online Supplements of this dissertation, we provide a first suggestion as to how insights from loss aversion may be included in economic evaluations, by incorporating a *loss aversion premium*. When and why policy makers should include a loss aversion premium in economic evaluations, may be explored in future work taking a broad view of relevant factors in the decision making process. For simplicity, this approach, that involves deliberately adjusting the value assigned to changes between health states that involve losses, was applied with current health as reference point. Such a loss aversion premium could be used when this is deemed relevant and normatively acceptable.

Conclusion: Research agenda and policy implications

Besides more discussion on corollaries of the corrective approach, such as the perfect health gap and a loss aversion premium, several steps can be outlined for future research. We suggest that these are necessary for successful potential application of the corrective approach in the policy context. First, the robustness and validity of PT parameters obtained through the corrective approach should be determined, both individually and combined, since differences were observed between studies using different methods (Attema et al., 2018a, Attema et al., 2013, Attema et al., 2016, Kemel and Paraschiv, 2018, and Chapters 3 and 7 of this dissertation). A head-to-head comparison of these methods could provide a more in-depth analysis of these differences and their impact on correction. Second, research could focus on replicating and extending earlier work on the corrective approach, preferably with a sample representative of the relevant population and test the validity of individually corrected QALY weights. Third, future research should aim to clarify the effect of PT on QALY weights elicited with discrete choice experiments (DCE), as these are employed more frequently in large scale valuation studies (e.g. Versteegh et al., 2016). Given that orthogonal comparisons of TTO, SG and DCE are non-existent, only suggestive evidence exists showing that DCE weights are similar to classical TTO and SG weights (McCabe et al., 2006, Stolk et al., 2010). However, given that DCEs are typically applied assuming random utility models and since they use an aggregate approach to HSV, it may be difficult to reconcile with corrections at the individual level. Fourth, as mentioned, the corrective approach crucially depends on assumptions about the reference point. Future work should explore the role and nature of the reference point(s) further, especially for TTO (for example with an approach as in van Osch et al., 2006), and develop corrections for PT that are applicable when outcomes other than the

time spent in reduced health are taken as reference point. Finally, if the results of future research on correcting biases are encouraging, national tariffs using the corrective approach for the relevant health-utility classification, for example EQ-5D-5L or SF-6D, could be obtained to facilitate the incorporation of the corrective approach within health policy (as in Perpiñán et al., 2009).

Summarizing, if future research indeed demonstrates the merit of the corrective approach, our suggestion would be to apply the corrective approach in QALY measurement also in the context of actual decision making, which entails several steps:

1. In HSV exercises, for example in large scale valuation studies, measure each subject's degree of deviation from EU with the most accurate methods available and adjust individuals' responses accordingly. Although work exists that challenges some of its core presuppositions (e.g. suggesting no stable preferences exist at all: Slovic, 1995), PT appears to best capture these deviations.
2. If these corrected weights are found to be valid (and a better representation of health state preferences than classical weights), national tariffs could be calculated based on corrected weights. These could be used in economic evaluations informing policy makers.
3. Some of the correction factors used to 'clean' health state valuations, may still be informative for policy makers outside the context of HSV. We have explored how this may be true for loss aversion, in relation to the distinction between interventions producing health gains and those preventing health losses.

To conclude, despite developments and increased research efforts into the corrective approach, many unresolved issues still exist that caution against its widespread use. This suggests that the quest for improving methods for HSV, economic evaluations and decision making has clearly not ended yet. With this paper we hope to have encouraged both researchers and policy makers alike to explore these new opportunities.

9

Living up to expectations:
Experimental tests of
subjective life expectancy
as reference point in
time trade-off and
standard gamble

Chapter based on:

Lipman, S.A., Brouwer, W.B.F., & Attema, A.E. (2020). Living up to expectations: Experimental tests of subjective life expectancy as reference point in time trade-off and standard gamble. Journal of Health Economics, 102318

Abstract: Earlier work suggested that subjective life expectancy (SLE) functions as reference point in time trade-off (TTO), but has not tested or modeled this explicitly. In this paper we construct a model based on prospect theory to investigate these predictions more thoroughly. We report the first experimental test of reference-dependence with respect to SLE for TTO and extend this approach to standard gamble (SG). In two experiments, subjects' SLEs were used to construct different versions of 10-year TTO and SG tasks, with the gauge duration either described as occurring above or below life expectation. Our analyses suggest that both TTO and SG weights were affected by SLE as predicted by prospect theory with SLE as reference point. Subjects gave up fewer years in TTO and were less risk-tolerant in SG below SLE, implying that weights derived from these health state valuation methods for durations below SLE will be biased upwards.

Introduction

Time trade-off (TTO) and standard gamble (SG) are two popular methods to value health states, i.e. to obtain utility weights relevant for determining quality adjusted life-years (QALYs)³⁰. Although the methods share a similar purpose, their framing and outcomes differ substantially (Bleichrodt, 2002, Bleichrodt and Johannesson, 1997), with SG weights typically being higher than TTO weights (e.g. Read et al., 1984, Torrance, 1976). Bleichrodt (2002) proposed that these differences could be explained by differences in the theoretical assumptions underlying TTO and SG. Both methods' QALY weights are typically calculated using theoretical models that have been shown to be empirically invalid, i.e. expected utility theory (for violations, see: Llewellyn-Thomas et al., 1982, Starmer, 2000) and the linear QALY model (for violations, see: Abellan-Perpinan et al., 2006, Bleichrodt and Pinto, 2005). More specifically, TTO and SG weights are biased, according to Bleichrodt (2002), because individuals show several empirical deviations from these simplified models, including loss aversion, probability weighting, utility curvature and scale compatibility. The first three of these deviations can be modeled through prospect theory, and Bleichrodt (2002) proposed that such modelling could reduce the difference between TTO and SG, for which some empirical support was found in Chapter 7 of this dissertation.

Prospect theory was originally developed as an alternative to expected utility (EU) theory for decision making under risk and uncertainty (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992). Most importantly, prospect theory assumes reference-dependence, i.e. outcomes are not evaluated in final terms, but as changes relative to a reference point (RP). The RP is a neutral outcome, such as the status quo (i.e. current health), but many alternative comparators have been argued to be able to serve 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 or formed by aspirations, expectations, norms, and social comparisons (Tversky and Kahneman, 1991). It is, however, paramount to determine the exact location of the RP, as this 'neutral outcome' will divide all other outcomes into gains and losses. Within prospect theory, this is especially relevant, as it assumes loss aversion, i.e. losses (relative to the RP) carry more weight than gains of the same size. Furthermore, in prospect theory probabilities can be transformed non-linearly by means of a probability weighting function, which may also differ between gains and losses.

Often it remains unclear exactly how RPs are selected, and how RP selection should be modeled within prospect theory (Wakker, 2010). Two different streams of literature have produced insights on the role of RPs in health-related decision making, which we try to unify in this paper. First, in applications of prospect theory to health outcomes, typically some plausible assumption is made about which outcome could serve as RP. This approach, where typically the RP is selected from the outcomes available *within* the scenarios presented to respondents, allows for tractable modelling and the formation of empirical predictions based on these assumptions. For example, earlier work on RP location for TTO and SG has suggested that the certain outcome, i.e. the impaired health state, will likely serve as RP in these health state valuation exercises (van Osch et al., 2006). Using such an approach to RP

³⁰ These weights are sometimes referred to as 'utilities'. We will use the term QALY weights, and TTO or SG weights to refer to QALY weights elicited by TTO and SG respectively.

selection, prospect theory has been successfully applied in the health domain, for example, earlier work showed that the main tenets of prospect theory (e.g. loss aversion and probability weighting) apply to decisions about human lives (Kemel and Paraschiv, 2018), length of life (Attema et al., 2013, Treadwell and Lenert, 1999, Verhoef et al., 1994, and Chapters 3 and 7 of this dissertation), and quality of life (Attema et al., 2016). Second, a literature exists suggesting that RPs that originate *outside* the specific decision task at hand may also be selected by respondents. Such studies typically observe some effect of these reference-outcomes on decision-making or well-being, and conjecture that this effect may be due to reference-dependence. Examples of such suggested reference-dependence are RPs based on expectations for length (van Nooten and Brouwer, 2004, van Nooten et al., 2009) and quality of life (Brouwer et al., 2005, Wouters et al., 2015), or social comparisons (Wouters, 2016).

In this paper, we focus on the effects of individuals' subjective life expectancy (SLE), i.e. self-reported anticipated length of life, which could serve as an RP as defined within prospect theory in health state valuations. It is well-known that many individuals expect to live longer than actuarial life expectancy (Brouwer and van Exel, 2005, Péntek et al., 2014, Rappange et al., 2016). Gauge durations in TTO typically do not coincide with these expectations; frequently, projected life span in TTO is considerably shorter. Although individuals may not be fully aware of this reduction during health state valuation (van Nooten et al., 2014), earlier work on SLE has consistently found that individuals with higher SLE gave up fewer years in TTO, and thus associated QALY weights were higher. We will refer to these changes in QALY weights for durations further away from expectations about length of life as the '*SLE effect*'. This SLE effect was found for 10-year TTOs (van Nooten et al., 2009), for patients valuing their own health state (Heintz et al., 2013), and for TTOs using a lifetime time-horizon (van Nooten and Brouwer, 2004). Considering that in most cases life years traded off in TTO fall short of SLE, it may seem plausible to assume that these life years are perceived as being in the loss domain already, and thus given up reluctantly (van Nooten and Brouwer, 2004, van Nooten et al., 2009). This earlier work postulated that the SLE effect may occur as a result of loss aversion, yielding unwillingness in TTO exercises to further reduce lifetime compared to individuals' SLE which serves as RP.

However, this explanation of the SLE effect has never been modeled adequately or tested directly, as earlier work on the SLE effect has relied on investigation of heterogeneity in SLE by means of an observational between-subjects approach, i.e. explaining differences in TTO responses by differences in SLE. Furthermore, if the SLE effect applies to TTO weights as a result of reference-dependence, one could also expect such effects on SG, as this method may also be affected by loss aversion (Bleichrodt, 2002). This has not yet been tested to our knowledge. Therefore, in this paper we extend earlier work on SLE effects in health state valuation by:

- i) Constructing a model based on prospect theory with reference-dependence with respect to SLE.
- ii) Reporting an elaborate test of the SLE effect by experimentally varying the tradable life years above and below SLE for both TTO and SG.

More specifically, by developing a model based on prospect theory we are able to construct tractable predictions about the SLE effect for TTO and SG responses, if it indeed serves as RP. We test these predictions by means of within-subjects experimental methodology in which we construct different versions of TTO and SG, to directly compare QALY weights

for life years both under and above SLE for each individual. Through this procedure we test if QALY weights differ for durations that can be perceived as either gains or losses compared to SLE and, hence, whether SLE functions as a formal reference point. This approach is applied in two experiments (labeled Study 1 and Study 2).

The remainder of this paper is structured as follows. First, we define our theoretical model based on prospect theory and, next, we derive predictions. Study 1 and Study 2 are reported in separate sections. Study 1 applied the experimental methodology with a convenience sample of students. The results of this study suggest that SLE indeed serves as RP for TTO and SG. In Study 2, the external validity of these findings is tested by recruiting a sample of individuals aged 60 years and older, largely confirming the results from Study 1. In the final sections we discuss these results and conclude.

Theoretical framework

Notation

TTO and SG are denoted as health profiles described as (Q, t) , where Q represents health status and t denotes the age at which the profile ends (e.g. living in a wheelchair until age 85), with D and FH denoting the states Dead and Full Health, respectively. Subscripts (e.g. a, r, x, y) are used to indicate chronic health profiles faced by a decision-maker with age t_a , where duration is defined as $T_x = t_x - t_a$. Importantly, t_a can, but need not, be the decision maker's current age (it could be any $t_a > 0$). Risky prospects are defined as $(Q_x, T_x)_p(Q_y, T_y)$, i.e. health profile (Q_x, T_x) with probability p , and health profile (Q_y, T_y) with probability $1 - p$. Preference relations are defined as usual, i.e. they are weak-ordered (complete and transitive), and denoted by $>$ (strict preference), \succsim (weak preference), and \sim (indifference).

The TTO method asks for a time equivalent in perfect health which yields indifference between T_x years in health state Q and T_y years in FH. The number of years in T_y is varied until the respondent is indifferent between the two options, i.e. $(Q_x, T_x) \sim (FH, T_y)$. The SG method involves a choice between a number of years (T_x) in health state Q_x for certain and a gamble with two outcomes, which are FH during the same time period (T_x) and D . Probability p is varied until the respondent is indifferent between the two alternatives, i.e. $(Q_x, T_x) \sim (FH, T_x)_p(D)$. Typically, preferences in TTO and SG are modeled within the general QALY model (Miyamoto and Eraker, 1989), which assumes that chronic health profiles (Q_x, T_x) can be evaluated by the utility function $V(\cdot)$:

$$V(Q_x, T_x) = U(Q_x) * L(T_x), \tag{9.1}$$

with $U(Q)$ denoting utility of health status and $L(T)$ denoting the utility of T life years.

Assuming $L(T) = T$ (i.e. the linear QALY model³¹), with the common normalization such that $U(FH) = 1$, TTO indifferences can be evaluated by:

$$U(Q_x) = \frac{T_y}{T_x}. \tag{9.2}$$

³¹ Note that this framework assumes no discounting of future life years, i.e. linear utility. This framework has been generalized to include non-linear utility by Miyamoto and Eraker (1989).

SG indifference, on the other hand, additionally assuming EU and $V(D) = 0$, can be evaluated by:

$$U(Q_x) = p. \quad (9.3)$$

Although Eq. 9.2 and Eq. 9.3 are only valid under these strict assumptions (more general derivations are available in Chapter 7), these equations are often used in large scale health state valuations (Brazier et al., 2002, Versteegh et al., 2016).

Reference-dependence model for SG and TTO

Reference points play no role within the frameworks of EU and the general QALY model. Thus, in order to test whether SLE serves as RP, we will supplement the generalized QALY model with prospect theory, following closely the model developed in Chapter 7 of this dissertation. This means that we assume that the general QALY model holds with the additional assumptions outlined below included.

We assume separate evaluations of gains and losses in life duration compared to an RP, denoted T_r . This RP is an expected health profile, which is taken to last for T_r years, starting from the age (t_a) of the decision maker until their SLE (t_r), i.e. $T_r = [t_a, t_r] = t_r - t_a$. Throughout we will denote durations of health profiles (Q_x, T_x) as deviations with respect to this RP as follows: we will write (Q_x, T_x^*) with $T_x^* = T_x - T_r$. For example, imagine a 50-year old subject with SLE of living until 80. The health profile of living in a wheelchair until age 70 will be denoted as (*living in wheelchair*, T_x^*) with $T_x^* = (t_x - t_a) - (t_r - t_a) = (70 - 50) - (80 - 50) = -10$ (for more examples, see the Online Supplements of this dissertation). We restrict our prospect theory model to life duration, even though it has been suggested that reference-dependence may also exist for health status (Brouwer et al., 2005, Wouters et al., 2015). However, both from a theoretical and from an empirical point of view such reference-dependence for Q is hard to approximate. That is, prospect theory is typically applied to single-attribute outcomes, such as money, while health profiles consist of both life duration and health status. Multi-attribute characterizations of prospect theory exist, but because health status is a qualitative measure, loss aversion is not theoretically meaningful for this attribute (Bleichrodt and Miyamoto, 2003).

As a solution, we apply an attribute-specific evaluation (Bleichrodt et al., 2009) by making three modifications to the general QALY model, to allow testing for reference-dependence with SLE as RP. First, we modify $L(T)$ in the general QALY model to $L^i(T^*)$, which is a standard ratio scale utility function, that can differ between gain outcomes (i.e. (Q_x, T_x^*) with $T_x^* \geq 0$, $i = +$) and loss outcomes (i.e. (Q_x, T_x^*) with $T_x^* < 0$, $i = -$), and is strictly increasing and real-valued. Second, loss aversion is incorporated into our model by taking $L^-(T^*) = \lambda L^+(T^*)$ for $T^* < 0$, where λ denotes a loss aversion index, with $\lambda > 1$ [$\lambda = 1$, $\lambda < 1$] indicating loss aversion [loss neutrality, gain seeking]. Third, we incorporate probability weighting, by evaluating probabilities in risky prospects by probability weighing functions w^i , $i = +, -$, that 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. We do not modify $U(Q)$ of the general QALY model, but we attempt to control for possible effects of reference-dependence of health status by applying our model only to health profiles where health status is better than what is considered acceptable at the ages under consideration. If, as Wouters et al. (2015) suggested, such acceptability serves as RP for health status, this restriction to

acceptable health states may avoid confounding effects as losses will only occur in terms of duration while health status will always be above expectation.

Thus, as in Chapter 7 of this dissertation, references over risky prospects with both gain and loss outcomes, i.e. $(Q_x, T_x^*)_p(Q_y, T_y^*)$, with $T_x^* \geq 0 > T_y^*$ are evaluated by:

$$w^+(p)U(Q_x)L^+(T_x^*) + w^-(1-p)U(Q_y)L^-(T_y^*), \quad (9.4)$$

while preferences over risky prospects $(Q_x, T_x^*)_p(Q_y, T_y^*)$ for either gains or losses are evaluated by:

$$w^i(p)U(Q_x)L^i(T_x^*) + (1-w^i(p))U(Q_y)L^i(T_y^*), i = +, - \quad (9.5)$$

where $i = + [-]$ when $T_x^*, T_y^* > [<] 0$, i.e. both outcomes are gains or losses. In Chapter 7 of this dissertation we show that when $w^i(p) = p, \lambda = 1$, and no distinction is made between gains and losses (i.e. no reference-dependence), this model reduces to the general QALY model.

Predictions

In this paper we consider two versions of TTO and SG. Typically, TTO and SG involve 10-year durations that start at current age. Instead, in this paper, we let the 10-year period in a reduced health state, which occurs in both TTO and SG, a) start at SLE, i.e. $t_a = t_r$ or b) end at each individual's SLE, i.e. $t_a = t_r - 10$. If SLE functions as RP, for a) the gauge duration occurs completely above SLE and thus always involves considerations in the gain domain (because $t_a = t_r$ gives $T_x^*, T_y^* > 0$). Similarly, for b) the gauge duration occurs completely below SLE and thus involves trade-offs in the loss domain (because $t_a = t_r - 10$ gives $T_x^*, T_y^* \leq 0$). Therefore, we label versions with gauge durations completely above SLE as gain versions (i.e. TTO-gains and SG-gains), while those versions with life years occurring completely below SLE are labeled as loss versions (i.e. TTO-losses and SG-losses). To distinguish between the starting ages in these versions for gains and losses, we add superscripts g and l , i.e. $t_a^g = t_r$ and $t_a^l = t_r - 10$. As a final notational convention, given that both versions have the same durations T_x (10 years starting at different ages), for clarity, we will add superscripts to health status for health profiles (Q_x, T_x^*) , such that (Q_x^g, T_x^*) and (Q_x^l, T_x^*) refer to profiles in gain (starting at t_a^g) or loss versions (starting at t_a^l), respectively. For example, consider a subject expecting to live until age 80 ($t_r = 80$). She would receive gain versions with $t_a^g = 80$ and loss versions with $t_a^l = 70$. If SLE indeed serves as RP, this shift from t_a^g to t_a^l allows us to test the SLE effect, as it changes the perception of life years with respect to the RP.

In the remainder of this section, we will employ our theoretical model based on prospect theory with SLE as RP to derive predictions about the SLE effect on TTO and SG. We will obtain these predictions by illustrating the implications of our prospect theory model as opposed to a reference case, in which linear QALYs and EU hold (i.e. Eq. 9.2 and Eq. 9.3 can be applied). For the sake of brevity and clarity, we focus on providing graphical illustrations of these predictions in Figure 9.1 and Figure 9.2. A complete and formal proof of these predictions can be found in the Online Supplements of this dissertation

SLE effects for TTO

For TTO, consider as reference case, a subject willing to give up 2 years with reduced health status (Q_x) to obtain full health for 8 more years in the gain version. Using our notation with SLE as RP, this yields the following indifference: $(Q_x^g, 10) \sim (FH^g, 8)$. That is, in the gain version the subject is indifferent between gaining 10 years beyond SLE in health state Q_x and gaining 8 years in full health. We will derive predictions from our model as to what this indifference implies for the years given up in loss versions, i.e. predict T_y^* in $(Q_x^l, 0) \sim (FH^l, T_y^*)$. We first consider the reference case, with linear utility (i.e. $L^-(T^*) = L^+(T^*) = T^*$) and no loss aversion ($\lambda = 1$), which yields the following indifferences: $(Q_x^g, 10) \sim (FH^g, 8)$ and $(Q_x^l, 0) \sim (FH^l, -2)$ for gain and loss versions, respectively (as each year has the same value). In Figure 9.1, we have represented such a combination of indifferences more generally. Initially for the reference case, we observe symmetric indifferences: $(Q_x^g, T_x^*) \sim (FH^g, T_y^*)$ and $(Q_x^l, T_x^*) \sim (FH^l, T_y^*)$. That is, shifting t_a^g to t_a^l , which in our experiments with 10 year durations gives $T_r = T_x^*$ for losses, does not affect preferences, as $(T_x^* - T_y^*)$ is equal between gains and losses. These indifferences indicate that in both scenarios each year given up in Q_x (i.e. $T_x^* - T_y^*$) exactly offsets an equal part of the value of the quality of life gained $\Delta(U(FH) - U(Q_x))$. However, such a combination of indifferences does not take into account any discrepancies between gains and losses. In Figure 9.1 we provide two illustrations of how the SLE effect for TTO responses due to: a) non-linear utility curvature, and b) loss aversion.

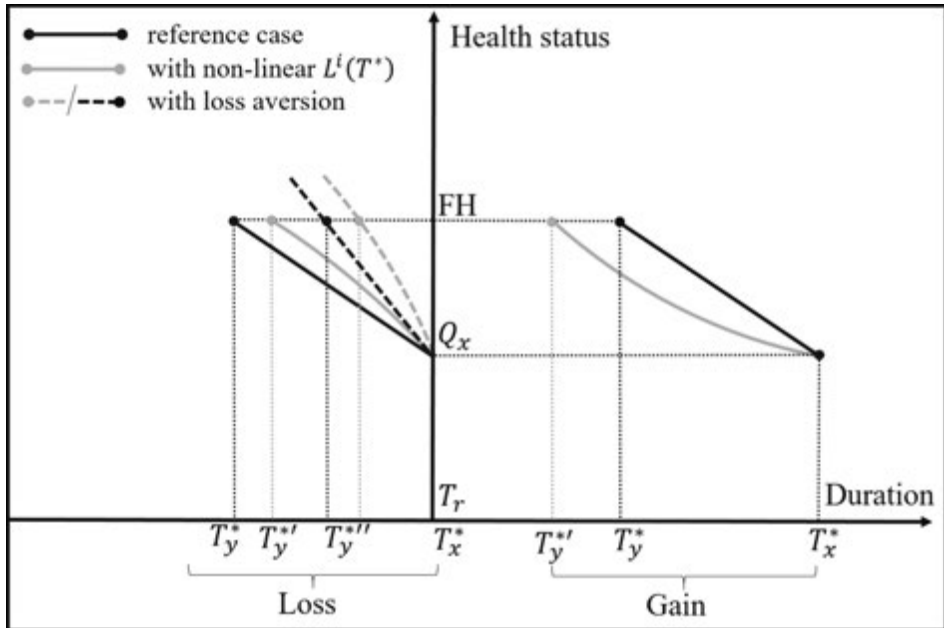


Figure 9.1. Indifference curves for time trade-off above and below SLE.

First, whereas TTO typically is derived assuming that utility of life duration is linear, i.e. $L^-(T^*) = L^+(T^*) = T^*$, earlier work on prospect theory has shown that this assumption is likely to be invalid for health outcomes (e.g. Attema et al., 2013, Kemel and Paraschiv, 2018, and Chapters 3 and 7 of this dissertation) and monetary outcomes (e.g. Abdellaoui, 2000, Abdellaoui et al., 2008, Abdellaoui et al., 2016, Bruhin et al., 2010). Instead, in prospect theory utility for gains is typically concave, and utility for losses is convex – i.e. utility for life duration is S-shaped. This inflection point in the utility curve may affect years given up in TTO-gains and TTO-losses versions, as it implies diminishing marginal sensitivity for additional life years gained or lost further away from T_r , as opposed to the linearity assumed in the reference case. Hence, it becomes important to consider where the life years given up in Q_x , and the years in which improved quality of life $\Delta(U(FH) - U(Q_x))$ is realized, fall along this S-shaped curve (we illustrate these effects in Figure 9.1). For TTO-gains, the years given up in Q_x (e.g. between 8 and 10) are further away from T_r than the years in which improved quality of life $\Delta(U(FH) - U(Q_x))$ is realized (e.g. between 0 and 8). Given that utility for gains is concave, in contrast to the reference case where each year is valued equally, we should find that each year given up in Q_x gets less weight than each year in which improved quality of life ($\Delta(U(FH) - U(Q_x))$) is experienced. Compared to the linear reference-case, this yields a convex indifference curve, and the respondent will give up more life years to offset the improvement in quality of life ($\Delta(U(FH) - U(Q_x))$) and restore indifference. Hence, we obtain $(Q_x^g, T_x^*) \sim (FH^g, T_y^{*'})$, with $T_y^{*'} < T_y^*$. For TTO-losses, however, the years given up in Q_x (e.g. between 0 and -2) occur closer to T_r than the years in which $\Delta(U(FH) - U(Q_x))$ is realized (between -10 and -2). As such, when utility for losses in life duration is convex, each year in which the improvement in quality of life is obtained gets less weight than each year given up. As a result, as compared to the reference case, this yields a concave indifference curve, and the respondent should give up fewer years to offset the improvement in quality of life ($\Delta(U(FH) - U(Q_x))$) and restore indifference. Hence, we obtain $(Q_x^l, T_x^*) \sim (FH^l, T_y^{*'})$, with $T_y^{*'} > T_y^*$.

Second, we take into account loss aversion, i.e. increased sensitivity to losses relative to T_r . Loss aversion yields reluctance to give up life years, and to account for this effect each year given up in Q_x should offset a larger part of the quality of life gained $\Delta(U(FH) - U(Q_x))$. This yields the steeper indifference curve in Figure 9.1, compared to the reference case where people are equally sensitive to gains and losses. As a result, if one is loss averse and durations in TTO occur below T_r , fewer years ($T_y^{*''}$) should be given up to restore indifference, yielding $(Q_x^l, T_x^*) \sim (FH^l, T_y^{*''})$. Thus, we predict that loss aversion with respect to SLE will decrease the years given up for TTO-losses versions as compared to gain versions. This conclusion also holds when taking into account non-linearity in the utility curve for life duration (see Figure 9.1).

SLE effects for SG

For SG, consider a subject willing to accept at most a 20% risk of immediate death for SG-gains. In our notation, this yields the following indifferences for gains:

$(Q_x^g, 10) \sim (FH^g, 10)_{0.8}(D)$. We will derive predictions from our model as to what this indifference implies for probability of death accepted in loss versions. In the reference case, linear QALYs and EU hold, i.e. the subject will also accept at most a risk of 20% of immediate death for the loss version, i.e. $(Q_x^l, 0) \sim (FH^l, 0)_{0.8}(D)$. In Figure 9.2, we have

represented such a combination of preferences more generally. Initially we observe the same indifference, i.e. $(Q_x^g, T_x^*) \sim (FH^g, T_x^*)_{p_g}(D)$ and $(Q_x^l, T_x^*) \sim (FH^l, T_x^*)_{p_l}(D)$ with $p_g = p_l$. That is, shifting t_a^g to t_a^l does not affect preferences, i.e. people are willing to risk the same probability of (D) for both SG-gains and SG-losses. This combination of indifferences in the reference case indicates that in both scenarios the possibility of an improvement in quality of life $\Delta(U(FH) - U(Q_x))$ for T_x exactly offsets the generally small chance of dying immediately. In case the difference in quality of life increases, i.e. when $(\Delta(U(FH) - U(Q_x)))$ increases, a larger chance of dying immediately will be accepted. However, just as for TTO, such a combination of indifferences does not take into account any discrepancies in the evaluation of gains and losses.

In Figure 9.2 we provide two illustrations of how SG responses are affected when SLE serves as RP: a) probability weighting (which may be different between gain and loss versions), and b) loss aversion. First, whereas in the reference case, probabilities are treated linearly (and thus also identically between gains and losses), our model based on prospect theory allows non-linear probability weighting. Importantly, it is typically observed that probability weighting is less pronounced for losses compared to gains, that is probability weighting is less inverse-S shaped, which has been found for health outcomes (e.g. Attema et al., 2013, Attema et al., 2016) and monetary outcomes (e.g. Abdellaoui, 2000, Kahneman and Tversky, 1979, Tversky and Kahneman, 1992). This implies that if SLE serves as RP, the same (small) probability of an extreme outcome receives more decision weight for SG-gains version compared to SG-losses versions. Inversely, when we observe a gain version indifference $(Q_x^g, T_x^*) \sim (FH^g, T_x^*)_{p_g}(D)$, then a higher probability of the extreme outcome (D) may be accepted in the equivalent loss version $(Q_x^l, T_x^*) \sim (FH^l, T_x^*)_{p_l}(D)$. For example, imagine a subject with $p_g = 20\%$, which implies that immediate death with decision weight $w^+(0.20)$ offsets³² the possible gain of quality of life for $\Delta(U(FH) - U(Q_x))$ with duration T_x^* . When we have $w^+(p_g) > w^-(p_l)$ for $p_g = p_l$, an increase in p_l to p_l' is required to restore indifference, i.e. for the disutility of (D) to offset $\Delta(U(FH) - U(Q_x))$.

Second, we take into account loss aversion by again assuming increased sensitivity to the possibility of losing compared to SLE, i.e. to durations below T_r . Importantly, the consequence of immediate death (D) differs between gain and loss versions; in the SG-gain version, entails a 20% chance of living up to SLE, while for loss versions dying immediately means living 10 years shorter than expected (i.e. a loss). Hence, SG-gain versions, in our experiment, involved no losses compared to T_r , and were not affected by loss aversion. Hence, if losses are incurred more reluctantly, smaller probabilities ($p_l'' < p_g$) of a loss are accepted for the same difference in quality of life $(\Delta(U(FH) - U(Q_x)))$. In Figure 9.2 we illustrate this by a steeper indifference curve.

³² Typically, in applications of prospect theory outcomes are rank-ordered, where in binary gambles such as SG, probability p is taken to reflect the probability of the extreme outcome in that domain. For the sake of clarity, in these illustrations we deviated from these conventions by taking p_g and p_l to refer to the chance of immediate death in both versions. The Online Supplements of this dissertation shows how the conventional notation can be applied without loss of generality.

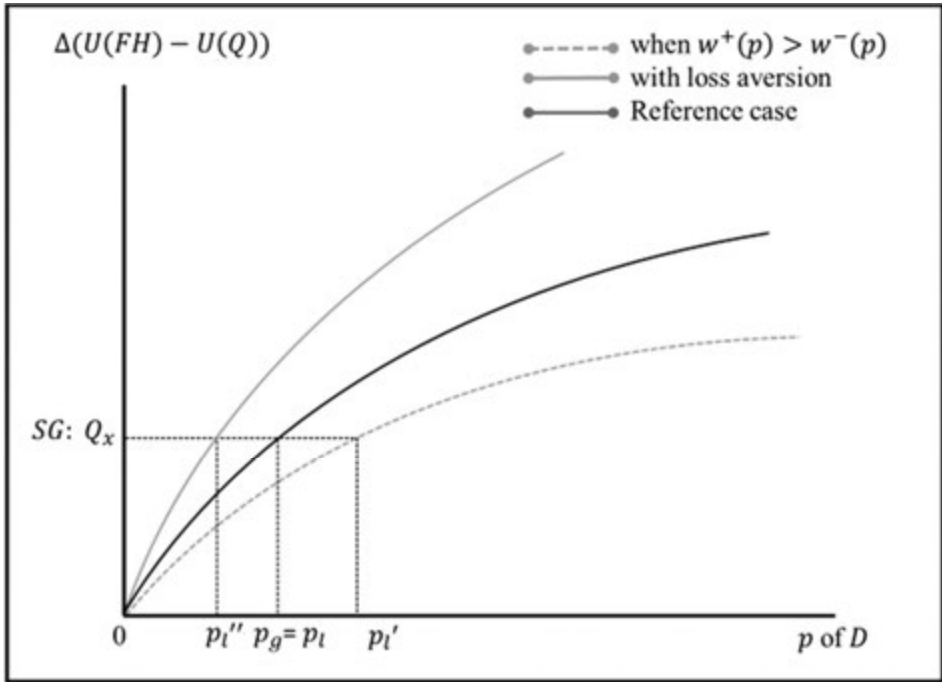


Figure 9.2. Indifference curves for standard gambles above and below SLE (superscripts refer to gains and losses).

Summarizing, for TTO our model predicts two SLE effects, both decreasing the life years given up for losses, while for SG our model predicts SLE effects in opposite directions, where the net direction is determined by the degree of loss aversion and differences in probability weighting for gains and losses. Given that these predictions differ between TTO and SG, shifting gauge duration from above to below SLE (i.e. moving from t_a^g to t_a^l) may yield different SLE effects between these two methods. We can derive no predictions about differences in magnitude of these SLE effects for TTO and SG, as they are affected by different components of prospect theory.

Study 1

In this first experiment, we tested our predicted SLE effects for TTO and SG in a lab experiment with a convenience sample of students.

Methods

This lab experiment started with several questions regarding expectations about length and quality of life followed by an elicitation of TTO and SG. Example instructions and screenshots can be found in the Online Supplements of this dissertation. The experiment used a 2 by 2 (method: TTO vs. SG, version: losses vs. gains) within-subjects design, with randomization by method. The experiment was completed by 102 Business Administration

students³³, recruited in the Erasmus Behavioral Lab in Rotterdam. A total of 71 males participated, and mean age for our sample was 20.25 (SD = 1.22). All students were rewarded course credit for participation in this 30-minute study.

Measures of expectations about length and quality of life

We measured students' expectations about length of life with the following questions (in this order): a) 'What is the minimum age you would hope to become?', b) 'What is the maximum age you would want to become?', and c) 'How old do you expect to become?'. The first two measures were obtained to explore how the typical estimates for SLE fall in between individuals' aspired minimum and maximum age, while question c) measures SLE, using a similar phrasing as van Nooten and et al. (2009). Students answered all three questions using a drop-down menu with answers in full years ranging from 30 to 120. To check if health states were considered acceptable, we also explored expectations about quality of life by obtaining a measure of acceptability for the health states that were used to apply TTO and SG (see Table 1). These questions were included as a manipulation check, to determine whether our model, which pertained to acceptable health states, can be applied. To introduce this concept, we used the following instruction (adapted from Wouters et al., 2015): 'In what follows you will receive questions regarding health at different ages. Generally, health deteriorates when we get older. Consider for example an 80 year-old person who is not able to walk further than 1 km. You might find this an acceptable condition for someone of 80, but less acceptable for 20 year old persons.' Next we asked them to rate all three health states using an identical drop-down menu ranging from 30 to 120, using the following question: 'Could you please indicate from which age onwards you find the following three health states acceptable?' Students could also answer 'Never', if they felt that a deteriorated health state was not acceptable at any age.

Operationalization of TTO and SG

All versions of TTO and SG featured a gauge duration of 10 years followed by immediate death, which is typical for this type of valuation exercises (van Nooten et al., 2009). These four valuation exercises (TTO-gains, TTO-losses, SG-gains and SG-losses) were all completed for three health states described by means of the EQ-5D-5L classification system (see Table 9.1 for the selected health states). TTO and SG were operationalized by using two-stage choice lists (see the Online Supplements of this dissertation), which were computerized via Qualtrics to prohibit multiple switching and violations of (stochastic) dominance within each choice list. For TTO, a first choice list identified indifference in years, and afterwards in months in a second choice list. For SG, choice lists elicited indifference with a first choice list identifying indifference at probability intervals of 10%, and afterwards specifying this in percentage points in a second choice list.

³³ Power analysis for paired t-tests with $n = 102$, the recommended power of 0.8 (Cohen, 1988) and a significance level of 0.05 indicated that it is adequately powered to detect differences with effect sizes as low as Cohen's $d=0.28$, indicating medium to small effect sizes (Cohen, 1988).

Table 9.1. Health states used in experiment

Dimension (EQ-5D)	Description	Best (Q1): 21211	Middle (Q2): 31221	Worst (Q3): 32331
Mobility	You have ... problems with walking	Slight	Moderate	Moderate
Self-care	You have ... problems with washing and dressing yourself	No	No	Slight
Usual activities	You have ... problems with your usual activities	Slight	Slight	Moderate
Pain/discomfort	You have ... pain or discomfort	No	Slight	Moderate
Anxiety/depression	You are ... anxious or depressed.	Not	Not	Not

Results

Table 9.2 reports descriptive statistics for our measures of expectations about length and quality of life. On average, students expected to become close to 85 years old, while wishing to become at least around 77 and at most close to 100 years. As we restricted our theoretical analyses to health states considered acceptable, we determined if students deemed health states Q1, Q2 and Q3 acceptable at all ages used in implemented TTO or SG versions. Overall, health states Q1 and Q2 were considered acceptable by most students, for all ages considered in this experiment, with 84% (Q1) and 72% (Q2) of our sample indicating that such a health status is acceptable from a lower age than the ages considered in our experiment. The most severe health state (Q3) was not considered acceptable at the lowest age considered (i.e. t_a^l), with only 34% of our sample considering such health problems acceptable at the ages presented in the loss versions of TTO and SG. For gain versions, this percentage was considerably higher, at 80%. This indicates that if reference-dependence exists for health status (as proposed by Wouters et al., 2015), this RP may fall in between the ages considered in the gain and loss versions of TTO and SG for health state Q3. However, we find relatively little non-trading (i.e. QALY weights of 1), with rates of non-trading from as low as 2% to 18% of the sample. Hence, to see if acceptability affected our main results, we ran several tests to explore whether this violation of the simplifying assumptions as described in our theoretical model affects TTO and SG responses (see the Online Supplements of this dissertation). We did not observe such an effect of acceptability of health status on TTO and SG responses. As such, we report our main results without excluding respondents from the sample.

Table 9.2. Medians, inter-quartile range (first quartile, third quartile), means and standard deviations for measured health outcomes for full sample ($n = 102$)

Outcomes	Median	IQR	Mean	SD
SLE	85.0	(80.00, 90.00)	84.68	9.56
SLE-min	80.0	(70.00, 85.00)	77.20	11.42
SLE-max	100.0	(93.00, 105.00)	99.91	11.80
Acceptable age Q1	60.0	(55.00, 67.00)	59.55	11.23
Acceptable age Q2	70.0	(63.50, 75.00)	67.92	10.08
Acceptable age Q3	79.5	(70.00, 82.75)	76.35	9.71

Testing predicted SLE effects for TTO

First, we tested our predictions about SLE effects in the two versions of TTO (i.e. TTO-gains and TTO-losses). Table 9.3 shows aggregate results for TTO responses in both versions. In accordance with our predictions, fewer life years were given up in loss versions of TTO compared to gain versions for all health states (Wilcoxon tests, all p 's < 0.001). According to our model, this suggests that students would either be loss averse or showed less pronounced utility curvature for losses in life duration. Inversely, giving up fewer life years for loss versions will yield higher TTO weights, i.e. higher QALY weights assigned to the same health state. When we analysed our data at within-subjects, we observed that for Q1, Q2 and Q3 respectively, 61, 65, 68% of sample gave up fewer life years in loss-versions. For all three health states, these proportions were significantly larger than the part of our sample that gave up equal life years for both versions, or more life years for loss-versions (all χ^2 's ($2, N = 102$) > 39.71 , all p 's < 0.001).

Testing SLE effects for SG

Next, we compared the probabilities of immediate death risked in SG between the two versions (i.e. SG-gain and SG-losses). As can be seen from Table 9.3, lower probabilities of immediate death were risked for loss versions of SG compared to gain versions (Wilcoxon tests, all p 's < 0.001). According to our theoretical model, this implies that the effect of loss aversion was more pronounced than that of differences in probability weighting. Inversely, this leads to the conclusion that SG with durations below SLE will yield higher QALY weights for the same health state. When we analyzed our data within-subjects, we observed that for Q1, Q2 and Q3 respectively, 51, 49, 51% of our sample was willing to take a smaller risk of immediate death in loss versions. For all three health states, these proportions were significantly larger than the part of our sample that assigned equal probabilities to both versions, or was willing to risk a higher chance of immediate death for loss versions (all χ^2 's ($2, N = 102$) > 10.65 , all p 's < 0.005).

Table 9.3. Median years given up in TTO and probability of death risk in SG, including within-subject differences between gain and loss versions

	Gains		Losses		Diff.	
TTO						
Years given up	Median	IQR	Median	IQR	Median	IQR
Best (Q1)	4.00	(2.00, 5.50)	2.13	(0.17, 4.23)	1.08***	(0.00, 3.48)
Middle (Q2)	5.00	(2.94, 6.50)	3.17	(1.04, 5.00)	1.00***	(0.00, 3.00)
Worst (Q3)	6.00	(4.50, 8.00)	5.00	(3.00, 6.67)	1.23***	(0.00, 3.00)
SG						
Probability of D	Median	IQR	Median	IQR	Median	IQR
Best (Q1)	20.5	(5.75, 35.00)	14.50	(2.25, 30.00)	1.00***	(0.00, 9.75)
Middle (Q2)	25.00	(20.00, 40.00)	21.00	(10.00, 35.00)	0.00**	(0.00, 10.00)
Worst (Q3)	38.00	(24.00, 50.00)	31.50	(20.00, 45.75)	1.00***	(0.00, 11.00)

Note: *, **, *** indicate that differences between gain and loss version were significant at $p < 0.05$, $p < 0.01$, and $p < 0.001$, for the Wilcoxon signed-rank test.

Comparing SLE effect between TTO and SG weights

In order to compare SLE effects between TTO and SG we needed to normalize weights obtained by these two health state valuation methods to fit on the same scale. We will achieve this normalization by applying the derivation of TTO and SG weights under EU and the linear QALY framework (i.e. Eq. 9.2 and Eq. 9.3)³⁴. Although this is inconsistent with our theoretical model based on prospect theory, it is in line with how TTO and SG responses are typically transformed into QALY weights (see for example: Brazier et al., 2002, Versteegh et al., 2016). Hence, these comparisons may also illustrate the direction and magnitude of reference-dependence with respect to SLE when TTO and SG weights are obtained when this is not accounted for.

Figure 9.3 illustrates the aggregate results for our sample. Within versions (i.e. gains or losses), SG weights were significantly higher than TTO weights (Wilcoxon tests, all p 's < 0.037). When comparing within valuation methods (i.e. TTO or SG), QALY weights for health state valuation exercises involving losses produced significantly higher QALY weights, both for TTO and SG (Wilcoxon tests, all p 's < 0.002). For both methods, the differences between gain and loss versions were of similar magnitude for Q1, Q2 and Q3 (not significantly different, Wilcoxon tests, p 's > 0.52). These findings indicate that shifting gauge duration below SLE resulted in an average increase in TTO weights of between 0.15 and 0.23. For SG, a similar pattern was observed, with significant differences between gains and losses visible, where moving life years below SLE increased SG weights on average by 0.02

³⁴ Derivations of TTO and SG weights under PT are available (see Chapter 7 of this dissertation), but require assumptions about or measurements of $L^1(T)$, $w^1(p)$ and λ . This is beyond the scope of this paper.

to 0.12. These SLE effects were significantly larger than 0, and larger for TTO weights compared to SG weights across all three health states (Wilcoxon tests, p 's <0.002). We validated these SLE effects using a mixed effects regression, which also showed that our conclusions appear to be unaffected by acceptability of the health states Q1, Q2 and Q3 or gender (see the Online Supplements of this dissertation). Finally, we tested whether the typical difference between TTO and SG weights is affected by moving the gauge duration below SLE. To this end, for each subject, we calculated a difference score between TTO and SG per health state, with difference scores being obtained within versions (e.g. TTO-gains vs. SG-gains). This TTO-SG difference was smaller for losses compared to gains (Wilcoxon tests, all p 's <0.02), but differences remained significantly larger than 0 (Wilcoxon tests with $\eta = 0$, all p 's <0.04). Collectively, these findings suggest that moving the gauge duration below SLE increases QALY weights, with this SLE effect being larger for TTO than for SG.

Discussion

This section briefly discusses the results of Study 1 and the main limitations of this experiment that are remedied in Study 2. In accordance with our theoretical predictions, we observed a reduced willingness to give up life years in the TTO-loss version compared to the TTO-gain version (i.e. SLE-effect for TTO). For SG, similar to the TTO results, subjects were reluctant to risk losing life years when deciding about life years that fell short of their expectations (i.e. SLE-effect for SG). When comparing normalized TTO and SG weights (calculated in the common way, based on EU and the linear QALY model), we observed that QALY weights increased when gauge durations were moved below SLE, albeit to a larger extent for TTO. Hence, the difference between TTO and SG was smaller for loss versions. However, the QALY weights elicited for Q1, Q2, and Q3 were low, especially compared to earlier work on health state valuation with general population samples (Devlin et al., 2018, Versteegh et al., 2016). The results from Versteegh et al. (2016) allowed calculating a QALY weight for health states representative of the Dutch general public's valuation (i.e. tariffs). For example, Q1, Q2 and Q3 were assigned valuations of 0.88, 0.79 and 0.68, respectively, which is considerably larger than the valuations in Study 1 (especially those elicited with gain versions).

The low QALY weights elicited in Study 1 suggest that students were willing to give up large proportions of their remaining life or accept high risks of immediate death, just to avoid living in health states with relatively minor problems. At least two reasons can be provided to doubt the validity of such responses to TTO and SG. First, students were paid course credits for participation in this study. Generally, in behavioral experiments in health it is preferred to use financial incentives to motivate respondents to carefully consider their responses (Galizzi and Wiesen, 2018). As such, without an incentive to provide effort in our modified versions of TTO and SG, it could be hypothesized that students invested too little effort in considering the consequences of their choices. Hence, to resolve this issue, in Study 2 respondents were provided with a monetary reward for participation. Second, this first exploration of the process by which SLE affects valuations relied on a convenience sample of students. Obviously, this sample is small and not representative for the Dutch population in terms of age and education level. All students were required to imagine being much older than their current age and living in health states they were unlikely to have experienced. This could be problematic, as earlier work has shown that individuals may experience difficulty accurately predicting their future choices (i.e. projection bias, see: Loewenstein et al., 2003).

Furthermore, earlier work has found that SLE is associated with both age and education level (Brouwer and van Exel, 2005, Péntek et al., 2014, Rappange et al., 2016), and TTO depends on attitudes regarding ageing and end-of-life (Van Nooten et al., 2016), which may also be different for students compared to older populations. Hence, in Study 2 we applied our empirical tests in a sample of older persons to investigate the external validity of our findings

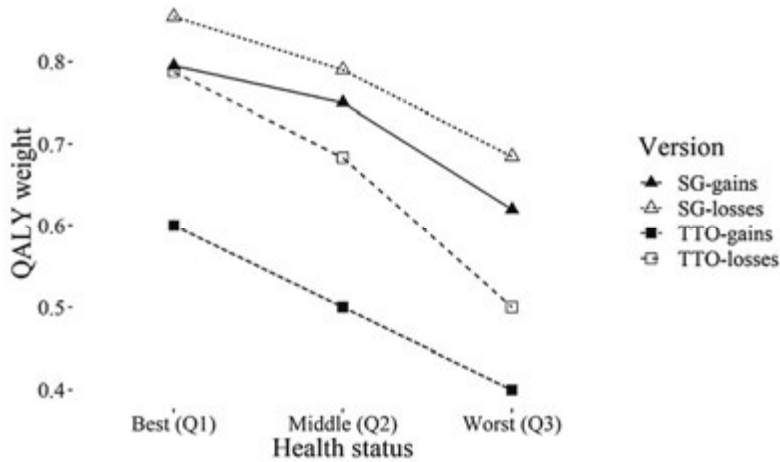


Figure 9.3. Median QALY weights for different versions and health states.

Study 2

In the second experiment, to test the external validity of our findings, we aimed to replicate our predicted SLE effects for TTO and SG in an online experiment with individuals aged 60 years and older. The methods were almost identical to that of Study 1, and as such we will only highlight modifications to the method below. Furthermore, seeing as we applied a similar analysis strategy, we will present the results of Study 2 without repeating the details.

Methods

Study 2 used the same measures of expectations about length and quality of life, health states and operationalizations of TTO and SG as were used in Study 1. However, the experimental task (programmed in Qualtrics) was now distributed online to a sample of 328 people aged 60 years and older. This was done through Prolific, a platform for online research with a large sample of individuals, who mostly live in the UK and US. It allows screening for a wide array of demographics, including age. When this experiment was run, Prolific had around 2600 users that were eligible (i.e. 60 years and older) and active in the last 90 days. Respondents were rewarded 3£ for taking part in this experiment. On average it took respondents 24 minutes to complete the experiment. Only a single question was added to the original set up in Study 1, i.e. a question to investigate experience with chronic illness.³⁵ Demographic characteristics for this sample of older people can be found in Table 9.4.

³⁵ 'Do you have a health condition or disease with long lasting effects (i.e. a chronic condition)? Examples of chronic conditions are: arthritis, COPD, asthma, diabetes, hepatitis, AIDS, and cancer.'

Table 9.4. Sample characteristics for sample older respondents (Study 2)

Demographic	Categories	Frequency	%
Age	60 – 65	202	61.6
	66 – 70	81	24.7
	70+	41	12.5
Sex	Male	127	38.7
	Female	201	61.3
Has chronic disease	Yes	120	36.6
	No	208	63.4
Nationality	United Kingdom	220	67.1
	United States	84	24.4
	Other	24	7.3
Relationship status	Married/in a relationship	187	57.0
	Divorced/Separated/Single	58	17.7
	Widowed	24	7.3
Highest education completed	Doctorate degree (e.g. Ph.D.)	6	1.8
	Graduate degree (e.g. M.Sc.)	39	11.9
	Undergraduate degree (e.g. B.Sc.)	93	28.4
	Technical/Community college	45	13.7
	High school diploma	77	23.5
	No formal diploma	6	1.8
Household income (GBP)	£10,000 - £29,999	127	38.7
	£30,000 - £49,999	68	20.7
	£50,000 - £69,999	34	10.4
	£70,000 - £99,999	13	4.0
	£100,000 or more	12	3.7

Results

Table 9.5 reports descriptive statistics on expectations about length and quality of life. The findings for SLE were similar to Study 1 with median SLE being 85 years old. Compared to the students in Study 1, respondents wished to become significantly less old (i.e. SLE-max was smaller) and considered impaired health states acceptable from a higher age onwards (Wilcox tests, p 's <0.001). Consequently, only 53% (Q1), 36% (Q2) and 16% (Q3) of the

respondents considered presented health states acceptable at all ages considered in the experiment. These were significantly smaller proportions than observed in Study 1 (all χ^2 's (2, $N = 328$) > 47.95 , all p 's < 0.001). A risk of having older respondents completing the experiment is that t_a^l (i.e. the age they are asked to imagine to be in the loss versions of TTO and SG) is lower than their current age. This was the case for 32 respondents (10% of the sample). However, excluding these respondents did not affect our results (see the Online Supplements of this dissertation). Furthermore, compared to Study 1 for all conditions and health states we found larger amounts of non-trading with rates of non-trading ranging from 12.5% to 34% of the sample (all χ^2 's (2, $N = 328$) > 3.93 , all p 's < 0.05). As for Study 1, several analyses were performed to check if acceptability of health states affected QALY weights or the main conclusions of our study (see the Online Supplements of this dissertation). We also included having a chronic disease in these analyses. Acceptability did not affect QALY weights, but experience with chronic disease was associated with higher QALY weights. However, our main results were similar for those with and without experience with disease (see the Online Supplements of this dissertation). Hence, we report on the full sample below.

Table 9.5. Medians, inter-quartile range (first quartile, third quartile), means and standard deviations for measured health outcomes for older sample ($n = 328$)

Outcomes	Median	IQR	Mean	SD
SLE	85.00	(80.00, 90.00)	84.68	9.56
SLE-min	84.00	(80.00, 86.00)	82.43	7.30
SLE-max	95.00	(90.00, 100.00)	93.93	9.38
Acceptable age Q1	70.00	(65.00, 80.00)	70.21	11.23
Acceptable age Q2	75.00	(70.50, 82.00)	75.50	10.52
Acceptable age Q3	80.00	(75.00, 86.25)	80.79	9.79

Testing predicted SLE effects for TTO and SG

Table 9.6 shows aggregate results for TTO and SG responses in both versions. As in Study 1, fewer life years were given up in the loss versions of TTO compared to gain versions for all health states (Wilcoxon tests, all p 's < 0.001). We observed that for Q1, Q2 and Q3 respectively 50%, 49%, and 41% of the sample gave up fewer life years in loss versions (rather than more or the same), which was a significant majority (all χ^2 's (2, $N = 328$) > 36.67 , all p 's < 0.001). As can be seen by comparing Tables 9.3 and 9.6, the SLE effect for TTO appears smaller for this older sample, but this difference was never significant (Wilcoxon test, all p 's > 0.06). In contrast to Study 1, we found no SLE-effect for SG, i.e. no evidence for lower probabilities of immediate death risked for the loss version compared to the gain version. We observed that for Q1, Q2 and Q3 respectively, 35, 30, 31% of our sample was willing to take a smaller risk of immediate death in loss versions (with similar proportions of the sample taking higher risks for loss versions). As can be seen by comparing Tables 9.3 and 9.6, the SLE effect for SG was smaller in Study 2, and this difference was

indeed significant for all three health states (Wilcoxon test, all p 's < 0.002). Finally, we explored whether excluding non-trading responses affected our findings for SLE effects for TTO and SG. Although this indeed increased effect sizes for TTO, the conclusions remained qualitatively similar (see the Online Supplements of this dissertation).

Table 9.6. Median years given up in TTO and probability of death risk in SG, including within-subject differences between gain and loss versions

	Gains		Losses		Diff.	
TTO						
Years given up	Median	IQR	Median	IQR	Median	IQR
Best (Q1)	2.92	(0.25, 5.31)	1.17	(0.00, 3.50)	0.00***	(0.00, 2.63)
Middle (Q2)	3.29	(1.00, 5.71)	2.00	(0.08, 4.40)	0.08***	(0.00, 2.00)
Worst (Q3)	4.92	(2.00, 6.65)	3.08	(0.92, 5.50)	0.08***	(0.00, 2.31a)
SG						
Probability of D	Median	IQR	Median	IQR	Median	IQR
Best (Q1)	10.00	(5.75, 35.00)	10.00	(2.25, 30.00)	0.00	(0.00, 5.00)
Middle (Q2)	13.00	(20.00, 40.00)	12.00	(1.00, 30.00)	0.00	(0.00, 2.00)
Worst (Q3)	22.00	(24.00, 50.00)	21.00	(3.00, 40.75)	0.00	(-1.00, 3.00)

Note: *, **, *** indicate that differences between gain and loss version were significant at $p < 0.05$, $p < 0.01$, and $p < 0.001$, for the Wilcoxon signed-rank test.

Comparing SLE effect between TTO and SG weights

Figure 9.4 illustrates the aggregate normalized QALY weights for each version. For each condition, QALY weights were significantly higher compared to Study 1 (Wilcoxon tests, p 's < 0.04), except for SG losses for Q1 (Wilcoxon test, $p = 0.11$). We also compared our results against the QALY weights calculated using the results by Devlin et al. (2018), which represent QALY weights for a sample representative of the UK (i.e. the country of residence for most of our sample). We found that the QALY weights elicited in Study 2 were significantly closer to the estimates by Devlin et al. (2018) than those elicited in Study 1 (Wilcoxon test, p 's < 0.04), except for SG-losses for Q1 (Wilcoxon test, $p = 0.11$). Still, our older persons sample reported QALY weights that were significantly different from the Devlin et al. (2018) estimates for all conditions (Wilcoxon tests, p 's < 0.01), except for SG-gains and losses for Q2 and SG-gains for Q3 (Wilcoxon tests, p 's > 0.16). Within versions (i.e. gains or losses), SG weights were significantly higher than TTO weights (Wilcoxon tests, all p 's < 0.007). QALY weights for health state valuation exercises involving losses (compared to gains) produced significantly higher QALY weights for TTO (Wilcoxon tests, all p 's < 0.002), but not for SG (Wilcoxon tests, all p 's > 0.09). Shifting gauge duration below SLE resulted in an average increase in TTO weights of between 0.10 and 0.11. For SG,

no such pattern was observed, where moving life years below SLE increased SG weights by 0.001 to 0.02. As in Study 1, the Online Supplements of this dissertation report a qualification of these findings by means of linear mixed-effects regression. The TTO-SG difference was smaller for losses compared to gains (Wilcoxon tests, all p 's < 0.001), but differences remained significantly larger than 0 for all health states (Wilcoxon tests with $\eta = 0$, all p 's < .007). Collectively, these findings suggest that moving gauge duration below SLE increases QALY weights for TTO, but not for SG (which leads to smaller TTO-SG differences for loss versions).

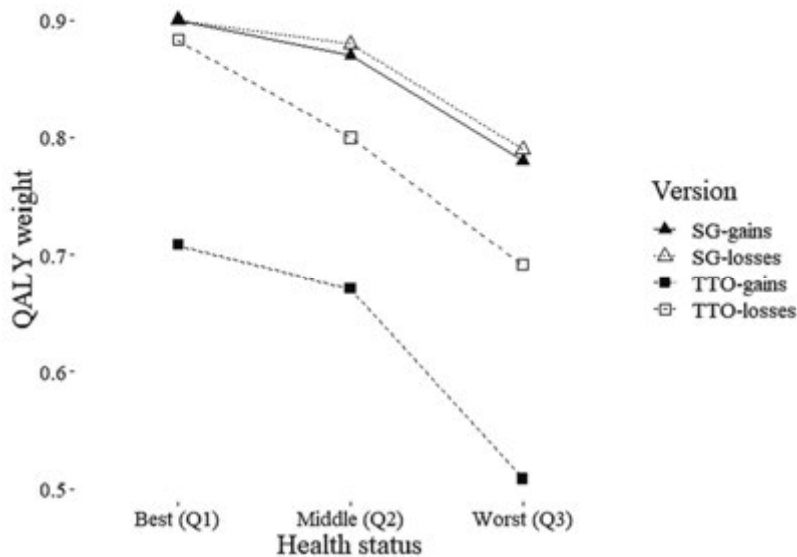


Figure 9.4. Median QALY weights for different versions and health states for older sample ($n = 328$)

General discussion

The goal of this paper was to (further) explore SLE effects for TTO and SG by means of a within-subjects approach. We constructed a theoretical model based on prospect theory, which allowed us to test its predictions using different versions for TTO and SG, with a gauge duration occurring either completely below (i.e. losses) or above SLE (i.e. gains). Although EU and the QALY model give no reason to expect differences between these TTO and SG versions, prospect theory, on the other hand, implies that if SLE functions as RP, loss aversion and sign-dependent evaluation of life years and probabilities can give rise to discrepancies between different versions. It was predicted that fewer years would be given up in TTO when elicited below SLE, i.e. TTO weights would be higher. Furthermore, for SG our predictions based on prospect theory suggest that, depending on their loss aversion and

probability weighting functions, individuals would be willing to either increase or decrease their risk of immediate death, i.e. the effect on SG weights is ambiguous.

We tested these predictions in two studies with a student (Study 1) and sample of individuals aged 60 years and older (Study 2). We find SLE to be similar to estimates from earlier work (Brouwer and van Exel, 2005, Péntek et al., 2014, Rappange et al., 2016, van Nooten et al., 2009). Furthermore, SLE falls in between maximum and minimum aspired ages, suggesting that it could indeed be taken as RP within prospect theory, as this is typically seen as a neutral position (Wakker, 2010). In accordance with our theoretical predictions, if life years in TTO occurred below SLE, we observed less willingness to give up life years in both Study 1 and Study 2. Hence, our results for TTO confirm the SLE effect observed in earlier work (Heintz et al., 2013, van Nooten and Brouwer, 2004, van Nooten et al., 2009, van Nooten et al., 2014) where similar comparisons were made between individuals expecting to live longer than TTO gauge duration or shorter. Furthermore, seeing as it occurs both in student and samples with older respondents, it appears to be robust to individuals' current age, which provides some support for the external validity of the effect of SLE on TTO. These findings (according to our model) suggest that: a) subjects refrain from giving up life years compared to SLE as a result of loss aversion (as suggested by van Nooten et al., 2009), and/or b) subjects show less diminishing marginal utility for life duration for losses compared to gains with respect to SLE.

For SG, the results for Study 1 were similar to those for TTO, i.e. students were more reluctant to risk losing life years when deciding about life years that fall short of their expectations. That is, when the gauge duration occurs below SLE, lower chances of immediate death were taken, which is, to our knowledge, a novel finding. However, these results were not replicated for people aged 60 years and older in Study 2. Given that our model based on prospect theory yields ambiguous predictions for SG, it can provide an explanation for this null result in Study 2. For Study 1, our findings suggest that loss aversion decreased willingness to risk immediate death for gauge durations below SLE. Our model predicts that probability weighting for gains and losses may have offset part of the effect due to loss aversion, which may explain the weaker effect of SLE for SG in Study 1, and perhaps the null result in Study 2. To explain the non-significant SLE effect for SG in Study 2 individuals aged 60 years and older should be less loss averse and/or had larger differences in probability weighting between gains and losses than the student sample in Study 1 had.

Although we derived predictions based on assumptions about loss aversion, utility curvature, and probability weighting, we did not include an empirical measurement of these prospect theory parameters. Instead, our predictions were based on earlier work on prospect theory for both health outcomes (e.g. Attema et al., 2018a, Attema et al., 2013, Attema et al., 2016, Kemel and Paraschiv, 2018, and Chapters 3 and 7 of this dissertation) and monetary outcomes (Abdellaoui et al., 2008, Bruhin et al., 2010, Kemel and Paraschiv, 2018), where substantial loss aversion and differences in curvature of probability weighting and/or utility functions between gains and losses were observed. As such, our study does not allow to directly test our theoretical explanations for the SLE effect for TTO, nor to determine why SLE effects could be observed in Study 1 but not in Study 2. Hence, combining our experimental approach with measurement of prospect theory parameters is a promising avenue for future research.

However, existing work suggests that older individuals are typically more loss averse and loss aversion decreases with education level (Arora and Kumari, 2015, Gächter et al., 2007), which would lead us to expect stronger loss aversion in the sample of Study 2. For probability weighting the evidence is inconclusive. Donkers et al. (2001) find some evidence that suggests more pronounced weighting of probabilities for higher ages, but they do not differentiate between gains and losses. As such, at least two alternative explanations for the null result for SG in Study 2 appear relevant. First, QALY weights for health states Q1, Q2 and Q3 were considerably higher in Study 2 compared to Study 1, with especially SG weights for these mild health states nearing 1.00. It may be possible that no SLE effect for SG is observed due to a ceiling effect, which could be tested by incorporating more severe states in future work. Second, it is possible that this null result is explained by differences between the student and older samples in how they perceive life (and death) at the ages considered in this experiment. Our results provide some indication for this, with students indicating to find health problems acceptable from younger ages than individuals aged 60 years and older. Future work could explore, for example using qualitative interview techniques, the influence of these perceptions on the effects of SLE on QALY weights.

Before arriving at the conclusion that SLE serves as RP in TTO and SG, several alternative explanations, not related to reference-dependence with respect to SLE, and methodological limitations should be considered. First, subjects in both studies were asked to imagine being older than their current age. If subjects did not adopt the instructions in our experiment, the gauge durations in this experiment would be strongly discounted (Attema and Brouwer, 2010b, van der Pol and Roux, 2005). Given that loss versions of TTO involved years below SLE, these would necessarily occur earlier in time than years given up in gain versions if current age is adopted instead of SLE. Thus, compared to their current age, life years in gain versions are likely to be discounted more strongly, and given up more willingly compared to life years for the loss version of TTO (i.e. this would predict higher QALY weights for loss versions). Similarly, if subjects used their current age instead of hypothetical ages in our experiment, the time dimension may explain higher utility for SG-losses compared to gains, as for monetary outcomes it is well-known (Abdellaoui et al., 2011, Baucells and Heukamp, 2012, Noussair and Wu, 2006) that risk-seeking increases when lotteries are resolved in the future. As such, SG-gains are resolved further away in the future than loss versions, and thus higher risk-seeking could explain the higher risks of death accepted for gain versions of SG. Hence, although it is not possible to make sure subjects indeed adopted our instructions, in the Online Supplements of this dissertation we show that if subjects did not adopt the ages in our experiment the effects of discounting would be negligible. Hence, given that we do find significant SLE effects, this is not likely to result from failure to adopt t_a^g and t_a^l .

Second, scale compatibility has been suggested to bias both SG and TTO (Bleichrodt, 2002, van Osch and Stiggelbout, 2008). Our manipulation, i.e. shifting life years around SLE, may have caused subjects to focus on life duration in TTO and SG. Given that life duration is fixed and equal in both options in SG, while in TTO life duration is varied along the choice list, this may explain the stronger effect of our manipulation on TTO, especially as the RP was also operationalized on the scale of life duration. Third, even though we provided respondents in Study 2 a monetary incentive to diligently complete our experiment, their rewards were not contingent on their choices (such incentive compatibility is typically preferred in economic experiments, Galizzi and Wiesen, 2018). Although earlier work in economics suggests that the use of hypothetical choices as opposed to incentive-compatible

choices has little to no effect on preferences (Camerer and Hogarth, 1999, Hertwig and Ortmann, 2001), we encourage the exploration of incentive-compatible choices in the context of health. Fourth, all theoretical predictions in this study were based on prospect theory, and hence, we explicitly assumed prospect theory to hold for decisions about health. Several authors have found violations of prospect theory, mostly for monetary outcomes (Bateman et al., 2007b, Birnbaum, 2006, Payne, 2005), but also for health (Feeny and Eng, 2005). As such, future work could explore if TTO and SG can be modeled in other reference-dependent models (Kőszegi and Rabin, 2006). Finally, to accommodate our subjects and avoid confusion or unnecessary errors, we maintained a consistent ordering throughout the experiment. Future work could explore whether this lack of counterbalancing between-subjects could have affected our conclusions, although other authors find no effects of order on gain-loss framing (e.g. De Dreu et al., 1994).

Conclusion

Whereas it is well-known that TTO and SG weights are typically different (e.g. Read et al., 1984, Torrance, 1976), earlier work on the role of SLE has exclusively focused on TTO. Our work suggests that decision-making in both health state valuation methods may be affected by subjective expectations about length of life, with QALY weights being higher for TTO and (to a lesser extent) SG when gauge durations are below SLE; i.e., SLE may serve as RP in health state valuation. This SLE effect could be relevant for the current practice in health state valuation, as this typically involves short gauge durations, which imply losses compared to their SLE for a large part of the sample. For example, when obtaining nationally representative TTO tariffs for EQ-5D, EuroQoL typically uses a 10 year duration for health states preferred to death (Oppe et al., 2014), which must fall short of SLE for many subjects. Applying derivations based on EU or linear QALYs will then yield TTO or SG weights that are too high.

Although finding a solution for this biasing effect attributable to SLE seems warranted, as discussed by Heintz et al. (2013), it can be complex to choose an appropriate duration for health state valuation. Durations below SLE may induce reluctance to lose any life years at all, while durations above SLE may yield lower QALY weights as individuals are more willing to lose some of these 'bonus years'. To our knowledge, no compelling normative argument exists to prefer either of these scenarios, suggesting that it may be necessary to acknowledge these possible biases and derive health state utility in a reference-dependent model (as discussed by: Abellan-Perpiñan et al., 2009, and in Chapters 7 and 8 of this dissertation). Therefore, we hope that our attempt to unify earlier work on reference points in health state valuation into a formal model based on prospect theory provides some insight into the consequences of not being able to live up to expectations about length of life.

10

A comparison of individual and collective decision making for standard gamble and time trade-off

Chapter based on:

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 tradeoff. The European Journal of Health Economics, 1-9

Abstract: Quality-Adjusted Life-Years (QALYs) are typically derived from individual preferences over health episodes. This paper reports the first experimental investigation into the effects of collective decision making on health valuations, using both time trade-off (TTO) and standard gamble (SG) tasks. We investigated collective decision making in dyads, by means of a mixed-subjects design where we control for learning effects. Our data suggest that collective decision making has little effect on decision quality, as no effects were observed on decision consistency and monotonicity for both methods. Furthermore, QALY weights remained similar between individual and collective decisions, and the typical difference in elicited weights between TTO and SG was not affected. These findings suggest that consulting with others has little effect on health state valuation, although learning may have. Additionally, our findings add to the literature of the effect of collective decision making, suggesting that no such effect occurs for TTO and SG.

Introduction

Many decisions about health are made in deliberation with others, e.g. children, spouses or medical professionals. This collective feature of decisions about health is, however, not typically reflected in health outcomes research focused on Quality-Adjusted Life-Years (QALYs). The weights representing quality of life, that are required to calculate these QALYs (i.e. QALY weights), are typically determined through choice-based methodologies (Dolan, 2000), such as standard gamble (SG) or time trade-off (TTO). Both methods are applied to the individual case, through decisions about one's own (hypothetical) health outcomes (Brazier et al., 2002, Devlin et al., 2018), i.e. no deliberation with others is allowed. As is well-documented in the health economic literature, QALY weights usually differ between SG and TTO (Bleichrodt and Johannesson, 1997, Read et al., 1984, Sackett and Torrance, 1978). SG weights are typically higher than TTO weights, and conventionally, this difference between SG and TTO was explained as resulting from deviations from the linear QALY model and expected utility (EU) theory which have both been found to be descriptively inaccurate (Abellan-Perpinan et al., 2006, Starmer, 2000, Wakker and Deneffe, 1996). Although it may be possible to measure these deviations and correct for their influences in SG and TTO (see Chapter 7 of this dissertation), currently no consensus exists on how these biases³⁶ are best measured or corrected for. Hence, the main motivation of this paper is to explore if the quality and outcomes of SG and TTO are affected by asking individuals to complete these tasks in groups, and if the difference between SG and TTO weights is reduced as a result.

The extant literature for monetary outcomes provides some indication that allowing individuals to discuss these complex decisions about health with others may be helpful. For example, collective decision making has been associated with less discounting and fewer time inconsistencies (Denant-Boemont et al., 2017). Other existing work on the effects of collective decision-making gives less firm results, with mixed evidence being reported for risk aversion (Ambrus et al., 2009, Brunette et al., 2015, Deck et al., 2012, Shupp and Williams, 2007, Zhang and Casari, 2012), ambiguity aversion (Brunette et al., 2015, Keck et al., 2014, Keller et al., 2007) and the violation rate of EU (Abdellaoui et al., 2013b, Bone et al., 1999, Rockenbach et al., 2007). When effects of collective decision making occur, they are hypothesized to result from the deliberation, bargaining and exchange of information when deciding collectively (e.g. Abdellaoui et al., 2013b, Deck et al., 2012). Taken together, these studies suggest that risk preferences, which are relevant for SG, and time preferences, which are relevant for TTO, might be affected by collective decision making. For example, discounting of future life years leads to downwards bias in TTO (Attema and Brouwer, 2014, Bleichrodt, 2002, van der Pol and Roux, 2005), and if such discounting is lower in when individuals decide in a group (Denant-Boemont et al., 2017) this could lead to higher TTO weights. Similarly, if groups are more willing to take risks (Brunette et al., 2015), perhaps due to reduced overweighting of small probabilities of dying, this could yield lower SG weights. If such effects occur simultaneously, the difference between SG and TTO might reduce.

³⁶ We will use the term 'biases' to refer to phenomena that yield violations of the linear QALY model and EU, the models that are used to calculate SG and TTO weights in practice.

Only a few studies exist documenting effects of deliberation in groups or deciding collectively on SG and TTO weights. McIntosh and colleagues (2007) found that completing SG in a panel and deliberating about responses decreased subsequent SG weights, and Karimi and colleagues (2019) found that deliberation in a panel had an effect on individual TTO weights. Just a single study explored collective valuation for both SG and TTO and found only small effects (Krabbe et al., 1996); however, this study used an anonymous voting system to obtain collective SG and TTO responses, i.e. deliberation between respondents was not allowed. Hence, those few studies on the effects of deliberation or collective decisions on QALY weights differ in several respects from the economic literature, in which typically smaller groups actually decide together (i.e. bargaining is included).

As such, we believe the evidence base on collective decision making precludes the formation of clear hypotheses for three reasons. First, next to the mixed evidence on risk preferences, an extensive psychological literature exists suggesting that in some cases detrimental effects of group decision making can be observed. This literature suggests that groups can engage in ‘groupthink’, which fosters limited information search and enhances confirmation bias (Esser, 1998, Janis, 1972). Second, the extant literature on collective decisions mostly studies monetary decision making, while SG and TTO involve health-related decision making, and these differ in many ways (Suter et al., 2015). Third, those few available investigations on effects of collective decisions for health (Akunne et al., 2006, Karimi et al., 2019, McIntosh et al., 2007, Robinson and Bryan, 2013) did not use an experimental design, i.e. often no control condition or comparator was in place. This complicates the interpretation of these studies’ findings, as these may be caused by learning instead (i.e. as a result of repeated measurement after deliberation). Indeed, it is well-known that such effects may occur in health state valuation (e.g. Augestad et al., 2012). Hence, in our work we explore the effects of collective decisions for SG and TTO, whilst controlling for learning effects.

Our study adds to the earlier literature on collective decisions and health state valuation in several respects. First, we report the first experimental test of the effects of collective decision making on QALY weights, by using a control condition constructed to control for learning. More specifically, we obtained a baseline measurement for SG and TTO for each subject, after which we distinguished between groups and individuals for repeated decisions. By using such a control condition (similar to that of Keck et al., 2014), we are able to isolate the effect of deciding collectively on multiple facets of SG and TTO decisions (only related to deliberation, bargaining and information exchange). We explore if such effects of collective decisions exist on internal consistency criteria, and if SG and TTO weights change by deciding collectively. Importantly, we test if the difference between SG and TTO reduces, as this could indicate that the different biases that are suggested to produce this difference are reduced (Bleichrodt, 2002). If that is the case, the use of collective decisions could provide an answer to the open questions surrounding the validity of QALY weights elicited with SG and TTO (see Chapter 8 of this dissertation). Finally, we test whether any possible effects of collective decision making carries over onto subsequent individual SG and TTO exercises for groups.

The remainder of the paper is organized as follows. We first cover the necessary notational conventions, introduce methodology and the experimental procedure. Next, the results are presented and the final section features a discussion of these results.

Preliminaries

In this paper, we only consider chronic health profiles described as (Q, T) , with Q denoting health status and T denoting its duration in years. For brevity, we denote immediate death as D and if health status is equal to full health (FH) we write $Q = FH$. Under the assumption of completeness, decision makers are able to form preferences over health profiles, denoted using the conventional notation: $>$, \succsim , and \sim to represent strict preference, weak preference, and indifference, respectively. Most studies applying SG or TTO assume that decision makers form these preferences as modeled within the linear QALY model³⁷ (Miyamoto and Eraker, 1989), i.e.:

$$V(Q, T) = U(Q) * T. \quad (10.1)$$

Decision makers decide about health profiles, either under certainty (in case of TTO) or under risk (in case of SG). Risk is operationalized by presenting decision maker with lotteries of the following form: $(Q, T)_p(Q', T')$, which signifies that health profile (Q, T) will be realized with probability p , and health profile (Q', T') with probability $1 - p$.

The SG method involves determining probability p at which decision makers are indifferent between a sure outcome (Q, T) , and a risky prospect $(FH, T)_p(D)$. Probability p is varied until the respondent is indifferent between a number of years (T) in health state Q for certain and a gamble with two outcomes, which are FH during the same time period (T), and D . These SG indifferences are typically evaluated under expected utility (EU) theory (Pliskin et al., 1980). The TTO method, on the other hand, asks for a time equivalent in perfect health which yields indifference between (Q, T) and (FH, T') , with $T > T'$. The number of years T' is varied until the respondent is indifferent between T years in health state Q and T' years in FH . Given the assumptions listed above, and setting $U(FH) = 1$ and $U(D) = 0$ the SG indifference $(Q, T) \sim (FH, T)_p(D)$ is evaluated by $U(Q) * T = p * (1 * T) + (1 - p) * 0$, and, thus: $U(Q) = p$. The TTO indifference $(Q, T) \sim (FH, T')$ is evaluated by: $U(Q) * T = 1 * T'$, and, thus, we obtain $U(Q) = T'/T$.

Bleichrodt (Bleichrodt, 2002) proposed that the typical differences between SG and TTO weights for the same health states may result from biases not accounted for in EU theory or the linear QALY framework, such as discounting, loss aversion and probability weighting. Thus, by evaluating SG and TTO without acknowledging these deviations, we should observe a gap between SG and TTO. Formally, we define the SG-TTO gap as the difference between $U(Q)$ as derived from SG and TTO: $\Delta(SG - TTO) = p - \left(\frac{T'}{T}\right)$. We expect $\Delta(SG - TTO) > 0$, and explore if collective decision making has effects on the difference between SG and TTO. If that is the case this gap should decrease, which we test empirically in an experiment.

³⁷ SG and TTO weights can be derived in more general models, which can account for some of the biases driving the differences between SG and TTO (see Chapter 7). However, such derivations are beyond the scope of this paper.

Experiment

Sample and design

A total of 163 Business Administration students (78 female, mean age = 19.37, $SD = 1.57$) participated in this experiment³⁸, which lasted around 55 minutes. Subjects were recruited via Erasmus Research Participation System, which rewards students with course credit for participation in scientific research. The experiment used a mixed between/within-subjects design (see Table 10.1) with two randomly assigned between-subjects conditions: individual decision making (IDM) and collective decision making (CDM). Experimental sessions were run on computers in sessions of two (CDM) or four (IDM) subjects sitting adjacently in separated cubicles. The experiment was programmed in Matlab, and instructions were provided on a separate sheet (see the Online Supplements of this dissertation). An instructor was present at all times to answer any questions subjects might have with regard to the procedure. Sessions consisted of three parts, with the experimental conditions IDM and CDM only differing in the second part. The first part served to establish a baseline measurement for SG and TTO weights, i.e. in Part 1 all subjects completed SG and TTO individually. In the second part, subjects in the CDM condition completed SG and TTO elicitation again collectively. Subjects in the IDM condition individually completed a filler task (adapted from Ameriks et al., 2007), which was not related to health states, risk or lotteries, to avoid confounding effects. The results of this filler task are not covered in this paper. In Part 3, to determine whether learning (IDM) or carryover effects (CDM) occurred, all subjects were presented with one final repetition of SG and TTO utility elicitation completed individually. When subjects finished Part 3, demographics were collected.

Measurements for SG and TTO

All SG and TTO elicitation were operationalized by using choice list methodology (see the Online Supplements of this dissertation for instructions and screenshots). This elicitation procedure, popularized by Holt and Laury (2002), is used frequently for elicitation of risk and time preferences for monetary outcomes (Andersen et al., 2006, Andersen et al., 2008). Although we are not aware of any study using choice list methodology for SG and TTO, recently choice lists have also been used to elicit preferences for health outcomes (e.g. Arrieta et al., 2017, Attema and Lipman, 2018). Figure 10.1 shows a combined example of SG and TTO choice lists.

For choice lists based on the SG method, subjects were faced with a choice between two alternatives. Alternative A would make them certain to live 50 more years³⁹ in some health state (Q), after which they would die. If they chose Alternative B, they would be taking a

³⁸ Sample size was informed by earlier studies on collective decision making in the economic literature. For example, the average sample size for all empirical studies on collective decision making cited in our Introduction is $n = 103$, with an average of $n = 40$ observations for groups (with sizes ranging from 2 to 5). Ethical approval was received from Erasmus Research Institute of Management's Internal Review Board, Section Experiments.

³⁹ Often, health state valuation studies use a 10-year duration (Oppe et al., 2014). For this student sample a 10-year duration followed by death would obviously entail a large decrease in life expectancy. It has been found that such a mismatch between durations in health state valuation and expectations about length of life may lead to biases in health state valuation (van Nooten & Brouwer, 2004; van Nooten et al., 2009; Chapter 9). Hence, we chose a much longer duration, more closely matched to our respondents' actual life expectancy.

gamble. The following instruction was used to clarify the risk of Alternative B: ‘On the one hand, you have the chance ($100 \times p\%$) of living 50 more years (T) in full health (i.e. no problems on any dimension), after which you will die, but on the other hand, you have a chance ($100 \times (1 - p)\%$) of dying within a week’. Subjects faced choice lists of 10 choices in which Alternative B varied; more specifically, p increased. For each elicitation, a two-pronged approach was used. First, p varied in increments of 10%, between 0% and 100%. After a switching point was obtained at this level, a second choice list was presented, which elicited a probability at the percentage point. For example, if a subject switched at $p = 80\%$ in the first choice list (as in Figure 10.1), she would face a second choice list that varied between 70% and 80% with increments of 1% (see the Online Supplements of this dissertation for screenshots).

Table 10.1. Overview experimental conditions

		Between-subjects comparisons	
Session		IDM ($n = 65$)	CDM ($n = 98$)
Within subjects	Part 1	Individual SG/TTO (I1)	Individual SG/TTO (I1)
	Part 2	Filler task (F)	Collective SG/TTO (G) *
	Part 3	Individual SG/TTO (I2) **	Individual SG/TTO (I2)

Note: * indicates the group effect, and ** indicates the carryover effect.

For TTO choice lists, Alternative A was the same as for the SG choice lists, i.e. living 50 more years (T) in the indicated health state (Q), followed by death. If subjects chose Alternative B, they would live T' more years in full health (i.e. no problems on any dimension), followed by death. A similar two-step elicitation procedure was in place, where, in the first choice list, T' varied between 0 and 50 years. In the second choice list, the indifference point of the first list was continued, and a more precise estimate was obtained by presenting subjects with a choice list with 10 increments of 0.5 year. For example, if a subject switched from A to B at $T' = 35$ years (as in Figure 10.1), she would face a choice list with Alternative B varying between 30 and 35 with 0.5 year increments (see the Online Supplements of this dissertation).

Collective decision making task

If a session was randomized to be a CDM session it consisted of two subjects who arrived at the lab at the same time. Both subjects first completed Part 1 individually, i.e. the baseline measurement for SG and TTO, while seated in separate cubicles. After they were both finished with Part 1 (if necessary one of the subjects was asked to wait until the other was finished), subjects were asked to move to one of the adjacent cubicles together. In this cubicle, they were asked to repeat the task they just performed (Part 2) and instructed to freely discuss amongst each other until they reached an answer that was satisfactory for both of them. Subjects, thus, filled out a single choice lists such as in Figure 10.1, which reflected

their joint evaluation. The experimenter remained present in the room during this time to address questions and monitor the experiment. The conversations between subjects were not recorded, we only store their collective response on the choice lists. After completing the collective task, subjects returned to their individual cubicle and were asked to complete Part 3 without discussing with each other.

SG		Common			TTO	
Alternative B	B	A	Alternative A	A	B	Alternative B
0% of 50 years in FH, D otherwise	<input type="checkbox"/>	<input checked="" type="checkbox"/>	50 years in Q	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0 years in FH
10% of 50 years in FH, D otherwise	<input type="checkbox"/>	<input checked="" type="checkbox"/>	50 years in Q	<input checked="" type="checkbox"/>	<input type="checkbox"/>	5 years in FH
20% of 50 years in FH, D otherwise	<input type="checkbox"/>	<input checked="" type="checkbox"/>	50 years in Q	<input checked="" type="checkbox"/>	<input type="checkbox"/>	10 years in FH
30% of 50 years in FH, D otherwise	<input type="checkbox"/>	<input checked="" type="checkbox"/>	50 years in Q	<input checked="" type="checkbox"/>	<input type="checkbox"/>	15 years in FH
40% of 50 years in FH, D otherwise	<input type="checkbox"/>	<input checked="" type="checkbox"/>	50 years in Q	<input checked="" type="checkbox"/>	<input type="checkbox"/>	20 years in FH
50% of 50 years in FH, D otherwise	<input type="checkbox"/>	<input checked="" type="checkbox"/>	50 years in Q	<input checked="" type="checkbox"/>	<input type="checkbox"/>	25 years in FH
60% of 50 years in FH, D otherwise	<input type="checkbox"/>	<input checked="" type="checkbox"/>	50 years in Q	<input checked="" type="checkbox"/>	<input type="checkbox"/>	30 years in FH
70% of 50 years in FH, D otherwise	<input type="checkbox"/>	<input checked="" type="checkbox"/>	50 years in Q	<input type="checkbox"/>	<input checked="" type="checkbox"/>	35 years in FH
80% of 50 years in FH, D otherwise	<input checked="" type="checkbox"/>	<input type="checkbox"/>	50 years in Q	<input type="checkbox"/>	<input checked="" type="checkbox"/>	40 years in FH
90% of 50 years in FH, D otherwise	<input checked="" type="checkbox"/>	<input type="checkbox"/>	50 years in Q	<input type="checkbox"/>	<input checked="" type="checkbox"/>	45 years in FH
100% of 50 years in FH, D otherwise	<input checked="" type="checkbox"/>	<input type="checkbox"/>	50 years in Q	<input type="checkbox"/>	<input checked="" type="checkbox"/>	50 years in FH

Figure 10.1. Example choice list for SG and TTO filled in by example participant. Note that FH and D denote health states full health and death respectively, ■ indicates that this choice is preferred by a hypothetical subject.

Health state descriptions

Each part consisted of SG and TTO elicitation for the same 3 health states (and 1 practice health state), for which descriptions were obtained from the EQ-5D-5L classification system (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”. 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). To familiarize subjects with the choice list elicitation, they completed a practice elicitation for Q_p : 41321

using both SG and TTO choice lists (in Parts 1 and 2). Next, SG and TTO elicitation were completed for three health states, which were relatively mild and ordered monotonically increasing in severity, i.e. each consecutive health state featured more severe problems on at least one domain and was identical otherwise. The following health states were used: 11221 ('high'), 21222 ('middle') and 32322 ('low'), which we denote Q_1 , Q_2 and Q_3 . We selected mild health states to avoid health states that may be considered worse than death, for practical reasons, as such severe health states require a different elicitation procedure (Oppe et al., 2014).

Data quality

Several checks for data quality were implemented. First, to familiarize subjects with health states Q_1 , Q_2 and Q_3 at the start of this experiment, subjects were required to rate these health states alongside death on a scale between 0 and 100, where 100 represented full health. Second, choice lists did not allow multiple switching points, which traditionally pose a significant problem to this method when it is applied with paper-and-pencil (Bruner, 2011). Third, to test for consistency, SG choice list elicitation were repeated for health state Q_1 in all Parts (before continuing with TTO). Finally, we were able to determine violations of monotonicity. Given that health states were monotonically increasing in severity, we should obtain $U(Q_1) > U(Q_2) > U(Q_3)$, i.e. monotonically increasing probabilities p accepted in SG and decreasing number of years T' in FH for TTO.

Analyses

We analyzed: a) decision quality, and b) decision outcomes (a full transcript of our analyses is available on request). Each of these decision domains was first analyzed by direct comparisons (i.e. t-tests) at the aggregate level between sessions and conditions. Second, we applied mixed effects regressions in order to i) determine if collective decisions for SG and TTO influences decision making beyond mere learning, and ii) estimate if collective decision making improves subsequent individual decision making. The former is referred to as a '*group effect*', while the latter is referred to as '*carryover effect*' (see Table 10.1). For the group effect we compared the group answers in the CDM condition (CDM: G) and the repeated individual answers in the control group (IDM: I2) to their respective baseline. Thus, this comparison consisted of the second time subjects completed SG and TTO weights for both conditions, while individuals in CDM completed this second round in groups. To estimate this group effect, we ran generalized linear mixed effect regressions (LMER) with subject random effects and the following fixed effects included: i) learning – dummy indicating whether it concerned a first or repeated session, ii) treatment – IDM or CDM, iii) method – SG or TTO and iv) group – interaction term for learning and treatment. The carryover effect was estimated similarly, where we instead compared CDM: I2 and IDM: I2 to their respective baseline. To estimate this carryover effect, we ran a similar LMER, with the same fixed effects included; i.e., i) learning, ii) treatment, iii) method, and iv) carryover–interaction term for learning and treatment. These analyses were performed with R using the lmerTest package. For the sake of brevity, we will not present full model statistics for the linear mixed-effect analyses, but only report fixed effect estimates (FEE) and standard errors (SE).

Results

Decision quality

We analyzed decision quality by determining the effect of collective decision making on our consistency checks and monotonicity of SG and TTO valuations (see the Online Supplements of this dissertation for additional results on decision quality).

Consistency

Consistency on repeated SG choices was adequate for all individual tasks (I1 and I2 for both IDM and CDM), with no significant difference between original and repeated elicitation (t-tests, p 's > 0.07). However, consistency was lower for collective decision making, with significant differences existing between original and repeated decision making (t-test, p < .001). Next, we estimated the group effect and carryover effect for consistency (see Table 10.2). Considering that consistency checks were only applied to SG, we dropped fixed effects for method in both analyses. We found no significant effects in mixed effects regressions.

Monotonicity

We determined for each subject to what extent violations of monotonicity occurred per session. A large majority (81% to 100% depending on session) of our subjects assigned monotonically decreasing QALY weights to all health states. Next, we estimated the group and carryover effect for monotonicity (see Table 10.2). Subjects were classified as either violators or non-violators; hence, we applied a linear binomial mixed effect model instead of LMER. First, when estimating the group effect, we observe significant effects for: a) treatment and b) group. This indicates that: a) although sampling was random, monotonicity was lower overall for subjects in CDM, and b) monotonicity increased for collective decisions above and beyond learning. No effects of learning or method were observed. Second, when estimating the carryover effect, we found no significant fixed effects.

Decision outcome

We analyzed decision outcomes using a similar analytical approach, with a focus on both absolute SG and TTO weights, and the relative differences between these methods.

SG and TTO weights

Figure 10.2 presents the main results on SG and TTO weights. Several trends at the aggregate level can be observed from this figure. First, QALY weights appeared to increase after repetition, with significant within-subjects increases for 9 out of 18 subsequent measurements (all p 's < 0.049). For example, for subjects in the IDM condition mean TTO and SG weights for Q_3 increased from 0.50 and 0.58 in I1 to 0.56 and 0.62 in I2, i.e. 0.06 and 0.04 respectively (see Figure 10.2 for the differences for all other measurements). Pooled across all measurements and subjects, each repeated measurement increased QALY weights by 0.03. Next, we estimated the carryover and group effect on the QALY weights, where we ran models with health state included as fixed effect. For both these approaches, we found a significant effect for a) learning, b) method and c) health state dummies. These effects indicate that a) repetition increases QALY weights, b) TTO weights were lower and c) the more severe health states received lower QALY weights. No effect of treatment, group or carryover was observed, indicating that the increase in QALY weights observed on aggregate appears not to be related to collective decisions.

Table 10.2. Fixed effect estimates (standard errors) for LMER analyses for both group and carryover effects

	Decision quality		Decision outcome	
	Consistency	Monotonicity ^a	QALY weight	$\Delta(\text{SG-TTO})$
Group effect : IDM: I1 vs. I2 CDM: I1 vs G				
Constant	8.87 (2.02) ***	1.09 (0.65) +	0.50 (0.03) ***	0.06 (0.02) ***
Learning	-1.68 (1.25)	0.64 (0.44)	0.04 (0.01) ***	0.00 (0.01)
Treatment: CDM	0.15 (2.59)	-2.75 (1.03) **	-0.03 (0.03)	0.02 (0.03)
Method: TTO		0.42 (0.28)	-0.03 (0.01) ***	
Group: (Learning*Treatment)	-0.74 (1.61)	2.38 (0.86) **	0.01 (0.01)	-0.01 (0.01)
Health state: middle			0.15 (0.01) ***	-0.04 (0.01) ***
Health state: high			0.29 (0.01) ***	-0.08 (0.01) ***
Carryover effect : IDM: I1 vs. I2 CDM: I1 vs I2				
Constant	8.87 (1.99) ***	1.32 (0.69) +	0.51 (0.02) ***	0.06 (0.02) *
Learning	-1.68 (1.22)	0.67 (0.45)	0.04 (0.01) ***	0.00 (0.01)
Treatment: CDM	-1.35 (2.30)	-0.51 (0.82)	-0.03 (0.03)	0.01 (0.03)
Method: TTO		0.42 (0.26)	-0.03 (0.01) ***	
Carryover (Learning*Treatment)	0.76 (1.31)	0.12 (0.55)	0.01 (0.01)	-0.00 (0.01)
Health state: middle			0.15 (0.01) ***	-0.03 (0.01) ***
Health state: high			0.28 (0.01) ***	-0.08 (0.01) ***

Note: *, **, and *** represent significance at $p < 0.05$, 0.01 and 0.001 respectively. ⁺ indicates marginal significance at $0.05 < p < 0.10$. ^a binomial regression.

Difference between SG and TTO

Next, to test if the difference between SG and TTO reduced as a result of collective decision making we compared the SG-TTO gap per session and health state (denoted $\Delta\text{SG-TTO}$). We found consistent evidence of higher weights for SG than TTO in health state Q_3 (paired t-tests, all p 's < 0.011). For example, for subjects in the IDM condition the mean SG-TTO gap was 0.08 for I1 and 0.06 for I2. However, we found no strong evidence for health state Q_2 (only significant for CDM-I2, paired t-test, $p < 0.01$) and Q_1 (paired t-tests, all p 's > 0.11). We observed a positive SG-TTO gap for baseline measurements (CDM/IDM-I1) pooled across health states with a size of 0.03 (significantly larger than 0, t-test, $p < 0.001$), suggesting that on average a difference existed between SG and TTO at baseline. Next, we applied our analytical approach to estimate group or carryover effects on this difference between SG and TTO (see Table 10.2). Only fixed effects for health states were significant, indicating that the difference between SG and TTO increased for more severe health states, and was unaffected by learning or collective decisions.

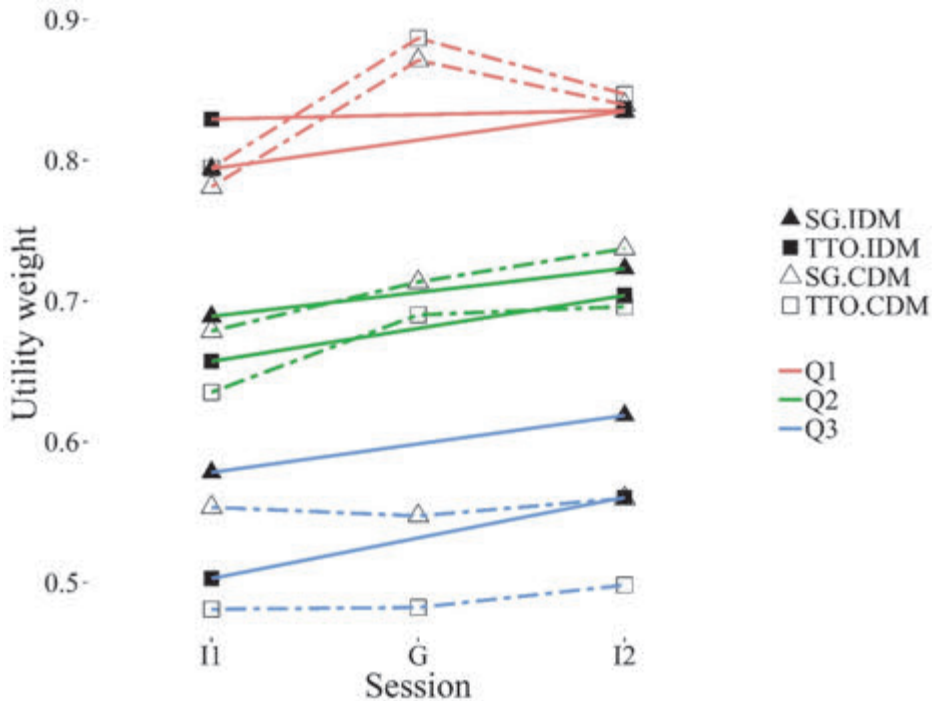


Figure 10.2. Mean weights split by method (SG vs. TTO), session (I1 vs. G vs. I2), health state (Q_1 vs. Q_2 vs. Q_3) and condition (IDM vs. CDM).

Discussion

In this study, we report the first experimental test of the effects of collective decision making on QALY weights. Collective decision making did not appear to have a systematic effect on quality of decisions for SG and TTO; no effects were found for consistency and the initially low monotonicity became up to par with individual decisions. Furthermore, we did find an effect of collective decision making with regard to outcomes for SG and TTO, although we found that QALY weights increased, both for collective decisions and for individual decisions. More sophisticated analyses indicated that this increase was only related to learning (and not to facets of collective decisions such as deliberation or bargaining), i.e. repetition increased SG and TTO weights (both in groups and individually). This trend of increased QALY weights for repetition is in accordance with earlier work by Augestad and colleagues (2012). Given that student samples in some cases have been found to yield low SG and TTO weights (e.g. see Chapter 7 of this dissertation) this learning effect could be seen as beneficial as it realized a movement towards QALY weights representative of the general population, obtained by a more comprehensive elicitation procedure, shorter SG and TTO durations, and a general public sample (Versteegh et al., 2016). As expected, we replicated the typical SG-TTO gap at baseline (Bleichrodt and Johannesson, 1997, Read et al., 1984, Sackett and Torrance, 1978), although this gap was less apparent for the least severe health state. Perhaps this lack of SG-TTO gap for the mildest health state results from a ceiling

effect (as QALY weights were close to 1.00). Importantly, we find that the SG-TTO gap was unaffected by collective decision making or learning, and no carryover effects were observed.

This study adds to the evidence base on collective decision making (mostly studies using monetary outcomes). In agreement with the mixed findings of those studies, we do not find a substantial beneficial effect of collective decisions for SG and TTO. However, earlier work on collective decisions for monetary choice suggested that groups discount the future less (Denant-Boemont et al., 2017). Because discounting has a negative effect on TTO values (Bleichrodt, 2002), less discounting in the group treatment would cause higher TTO values. Hence, our results could suggest that discounting of health outcomes is not affected by collective decision making; an alternative explanation would be that both discounting and loss aversion decrease in group tasks, which would neutralize each other (Bleichrodt, 2002). Our results also indicate that collective decision making does not alleviate the typical gap between SG and TTO, which is also partially explained as a result of discounting (Bleichrodt, 2002). Future research could therefore obtain separate measurements of discounting and loss aversion (and possibly also other traits such as scale compatibility and probability weighting) for health outcomes to test these possibilities.

Our results confirm findings by Krabbe and colleagues (1996), who find only small differences between collective and individual valuation for SG and TTO using an anonymous voting procedure, and that SG-TTO gaps are unaffected by collective decision making (Krabbe et al., 1996). Furthermore, our findings are in accordance with earlier non-experimental studies on deliberation in TTO or SG, which generally finds that deliberation has little to no effect on QALY weights (Karimi et al., 2019, McIntosh et al., 2007, Stein et al., 2006). Hence, it appears that deciding collectively has no added benefit beyond providing respondents with opportunities for learning. Our findings, thus, suggest that this procedure could be relevant for obtaining nationally representative value sets in settings where providing ample opportunity for learning is too costly or otherwise infeasible.

As in most experiments on the effects of collective decision making in the economic literature, the use of a student sample can be considered a drawback of this study. Obviously, students differ from the general population in a number of ways (so does any particular subsample used in empirical work). This is a common criticism of laboratory experiments, as any experiment using a non-representative sample will generate questions of external validity. To this end, in experimental economics, usually a distinction is made between experiments aimed at *measurement* and experiments aimed at documenting *treatment effects* (Jacquemet and L'Haridon, 2018). One can question the representativeness of our measurements, as individual QALY weights are typically lower compared to those in the general population (Versteegh et al., 2016). However, we believe it is not as straightforward to question external validity of the treatment effects (i.e. within-subject learning effects and between-subjects group effects), unless one explicitly assumes that these causal processes occur differently for our subject pool than for the general population. For one thing, students are likely to be younger, healthier and higher educated than the general population, and hence, the finding of a substantial learning effect in our student sample may suggest that the inclusion of a sufficient number of practice rounds in health state valuation will be necessary for a sample representative of the general public. We have no reason to expect that deliberation and bargaining are likely to occur differently for students as opposed to the general population. Nonetheless, a potential problem is that our sample consisted of students

exclusively, of whom it is likely that they had similar views on length and quality of life (and thus relatively few opportunities to influence each other). Hence, we believe future work should study if in less homogenous dyads (e.g. doctor/patient or student/retiree) the effects of collective decision-making on QALY weights are more prominent.

To conclude, our work suggests that collective decision making does not appear to yield an effect for health state valuation. As in earlier work (Krabbe et al., 1996), the difference between SG and TTO does not disappear when moving from an individual to a collective task, which suggests that collective decision making does not help to reduce the effect of the biases that affect SG and TTO (Bleichrodt, 2002). Therefore, other solutions for alleviating these confounding effects, such as more elaborate instructions, practice rounds and correction mechanisms (as developed in Chapter 7 of this dissertation) should be considered if one aims to correct for these biases.

11

CHAPTER 11

Discussion

Many individuals worldwide are living unhealthy lifestyles resulting in large preventable health losses. At the same time ageing populations and the desire to implement costly new health technologies (inter alia) are leading to growing pressure on health care budgets. These trends necessitate the study of *decisions about health*, both at an individual level and a societal level. Whereas research into individual decisions about health may provide insights that could be utilized to design interventions aimed at promoting a healthier lifestyle, the study of societal decisions about health is important in guiding how to face growing demand for health care in the context of a limited budget.

The first aim of this thesis was to improve our understanding of how individuals actually decide about their health. Rather than assuming that decisions about health are made rationally, i.e. maximizing expected utility (EU) by means of consistent and context-independent preferences, we studied health-related decision-making using methods and theories from behavioral economics (in which these assumptions are relaxed or dropped). As such, the studies in this dissertation drew from earlier work on financial decision-making showing that theories of rational decision-making provide a poor description of actual decisions. Instead, many individuals are influenced by reference-points (Baucells et al., 2011, Eibach and Ehrlinger, 2006, Koop and Johnson, 2012, Wang and Johnson, 2012), are disproportionately sensitive to losses (i.e. loss aversion, see: Abdellaoui et al., 2007, De Dreu et al., 1994, Gächter et al., 2007, Köbberling and Wakker, 2005, Tversky and Kahneman, 1991) overreact to small probabilities while underreacting to large probabilities (i.e. probability weighting, see: Abdellaoui, 2000, Bruhin et al., 2010, Fehr-Duda et al., 2006, Gonzalez and Wu, 1999, Suter et al., 2016), and are prone to be inconsistent (i.e. preference reversals, see: Butler and Loomes, 2007, Grether and Plott, 1979, Lichtenstein and Slovic, 1971, Reilly, 1982, Tversky et al., 1990). However, as such insights from behavioral and experimental economics are only limitedly applied in health economics (Galizzi and Wiesen, 2018), it is not entirely clear to what extent such concepts apply to decisions about health as well. Hence, in Part I of this dissertation a series of '*Behavioral experiments in health*' are reported, in which the following research questions were addressed:

1. To what extent are individuals risk averse for uncertain health outcomes with small or moderate stakes, and can EU explain such risk aversion? (Chapter 2)
2. Does the degree to which individuals are loss averse for life duration depend on the quality of life experienced during this time? (Chapter 3)
3. How heterogeneous are risk and time preferences, and can this heterogeneity be used to tailor financial incentives to improve decisions about health? (Chapter 4)
4. Are decisions about health as (in)consistent as those for money, and does the degree of preference reversals depend on who makes these decisions? (Chapter 5)

The second aim of this thesis was to use some of the insights derived from earlier work in behavioral economics to improve methods used in health state valuation. Currently, cost-utility analyses (CUAs) are performed with quality-adjusted life years (QALYs) derived by (implicitly) assuming that respondents completing health state valuation tasks have rational health preferences. More specifically, the methods used to derive the QALY weights (representing the utility of health status) are typically applied assuming the linear QALY model and expected utility (EU) theory. Instead, we acknowledge that many individuals deviate from the linear QALY model (and EU), and that such deviations may affect the outcomes of methods used to derive QALY weights. Earlier work has shown this is true

especially for two of the prominent methods used to derive QALY weights: time trade-off (TTO) and standard gamble (SG). Bleichrodt (2002) has suggested that the problematic difference between these methods might be explained by behavioral insights derived from prospect theory. However, empirical work had not yet fully tested this. Hence, Part II of this thesis reports several ‘*Applications of behavioral insights to health state valuation*’, which aimed to address the following research questions:

5. Which method reflects QALY weights better according to respondents themselves: TTO or SG? (Chapter 6)
6. Can prospect theory help to improve the validity of QALY weights (i.e. through correction)? (Chapter 7)
7. How feasible is it currently to apply ‘corrected QALY weights’ in practice? (Chapter 8)
8. What is the influence of subjective expectations of length of life on TTO and SG weights? (Chapter 9)
9. Can QALY weights be improved by deciding collectively during measurement? (Chapter 10)

Behavioral experiments in health

Chapters 2 and 3 reported behavioral experiments in health in which individuals decided about health gains and losses under risk. Although individual and cultural differences exist (Bontempo et al., 1997, Rosen et al., 2003, Weber and Hsee, 1998), many individuals dislike risks, and would rather have a certain outcome than a risky one with the same expected value. Such risk aversion is a hallmark phenomenon in health economics, for example in the classic work by Arrow (1963) on health insurance. Traditionally, decisions under risk are evaluated by expected utility (EU) theory with a concave utility function (von Neumann and Morgenstern, 1944). The degree to which EU is a descriptively valid model for decisions under risk, however, has long been questioned in economics, with many violations observed for financial decisions (Allais, 1953, Starmer, 2000). In Chapter 2, a now classic paradox, developed by Rabin (2000, 2001), was extended to the health domain in an experiment. Individuals were asked to report if they were willing to enter in a gamble for moderate stakes (‘50% chance of gaining 11 minutes of lifetime, or losing 10 minutes otherwise’) and a similar gamble with much larger stakes (‘gaining 400 days of lifetime with 50% chance or losing 4 days otherwise’). This experiment demonstrated that a large majority indeed show the ‘paradoxical’ preferences assumed by Rabin (2000, 2001), i.e. they turned down small stake gambles but were inclined to accept gambles for larger stakes (and thus violating EU). Hence, Chapter 2 answered research question 1, as the findings reported in this chapter showed that individuals are indeed risk averse for health outcomes, even for small stakes. Modelling this risk aversion by means of EU with concave utility over health outcomes, however, leads to implausible conclusions for the large stakes often present in the health domain. Hence, reference-dependent theories, which allow for sign-dependence and loss aversion should be considered for studying decisions about health.

Chapter 3 continued this investigation of decisions about health under risk, and as suggested in the previous chapter moved beyond EU by using prospect theory. Prospect theory is reference-dependent (as developed by: Kahneman and Tversky, 1979, Tversky and Kahneman, 1992, Wakker, 2010), as it assumes that outcomes are evaluated compared to a

reference-point, an outcome that serves to separate gains (outcomes above the reference-point) from losses (outcomes below the reference point). A key premise of prospect theory is loss aversion, i.e. the tendency for losses to loom larger than similarly sized gains, but only few studies exist that measure loss aversion for health outcomes. Earlier work that extended prospect theory to the health domain (e.g. Attema et al., 2013) required individuals to make decisions with different amounts of life duration as an outcome, but for simplicity the quality of life in which this life duration was spent was kept fixed (usually at perfect health). Hence, it was unclear if loss aversion depended on health status, or in other words, if the degree to which a loss of life duration looms larger than a similar gain in life duration depends on the quality of life experienced during this time. Therefore, in this chapter we tested whether loss aversion for life duration was independent of quality of life. Respondents in this study completed an adaptation of the non-parametric method developed by Abdellaoui et al. (2016) to measure loss aversion for different levels of quality of life. As such, they decided about health gains and losses, in which health was described as years of life in different health states. Separate loss aversion and utility curvature coefficients were estimated for each level of quality of life. Hence, Chapter 3 answered research question 2, as the findings of this chapter showed consistent loss aversion for life duration, independent of health status. The results of this chapter also suggest that while many people might be loss averse, the strength of this tendency may differ between contexts and individuals.

Similarly large heterogeneity can be observed in empirical work (e.g. reported in Chapter 8) that attempted to measure concepts such as: probability weighting (e.g. Abdellaoui, 2000, Bleichrodt and Pinto, 2000, Bleichrodt et al., 1999, Gonzalez and Wu, 1999, Suter et al., 2016), discounting (Attema et al., 2010, Attema and Brouwer, 2012a, Chapman, 1996, Chapman and Elstein, 1995, Loewenstein and Prelec, 1991, Redelmeier and Heller, 1993, Van der Pol and Cairns, 2000) and time inconsistency or present bias (Attema and Lipman, 2018, Attema et al., 2010, Bleichrodt et al., 2016, Rohde, 2010, Rohde, 2019, Shiba and Shimizu, 2019). This heterogeneity, however, is typically not considered when policy tools are used that are inspired by behavioral research. Instead, these policy tools often use a one-size-fits-all approach. In particular, they include some behavioral insight that is then assumed to apply to all people affected by the policy. For example, behavioral research on loss aversion has motivated offering financial incentives in which individuals could lose their own money (i.e. deposit contracts). Although such behaviorally inspired financial incentives have been found to be quite effective (Halpern et al., 2015, e.g. Kullgren et al., 2016), no conclusive evidence exists about which type of behaviorally inspired incentives work best, for whom and why (Adams et al., 2014, Haisley et al., 2012, Halpern et al., 2015, Paloyo et al., 2015). Also, the role that the large heterogeneity in risk and time preferences (such as loss aversion) play in preferences for behaviorally inspired incentives is unclear. Chapter 4 aimed to address this gap in the literature, by exploring the association between risk and time preferences and the type of behaviorally inspired incentives individuals would select into. More specifically, individuals were asked to tailor their own incentive scheme for reaching a weight loss goal, after which a wide array of risk and time preferences were elicited. The results in Chapter 4 allowed answering research question 3. The tailored incentives individuals designed for themselves were highly heterogeneous: e.g. a majority selected tailored incentives with a commitment component and only very few respondents preferred uncertain incentives. Nonetheless, these behaviorally inspired incentives were not related to risk and time preferences. Given that this null result has multiple possible explanations, it

remains unclear if tailoring incentives to individuals' preferences (e.g. measuring loss aversion beforehand and assigning those with high loss aversion a deposit contract) will improve effectiveness compared to one-size-fits-all behaviorally inspired incentives.

Throughout this dissertation, we have drawn on and expanded the work on financial decision-making, even though earlier work suggested that differences exist between decisions for money and health. For example, discounting is usually stronger for monetary outcomes when compared to health (Attema et al., 2018b, Chapman, 1996). Such differences between domains were also found for probability weighting (Suter et al., 2016) and loss aversion (Oliver, 2018). In Chapter 5, we extend this work by comparing choice consistency between domains, i.e. we study the degree of preference reversals in choices for both health and money. Unlike earlier work on preference reversals, Chapter 5 studied decisions made on behalf of others, which occur frequently in both domains. For example, health care professionals will make decisions on behalf of their patients, or at least provide them with advice on which treatment to choose. Similarly, financial professionals will decide how others' money is invested. Relatively few studies, however, have investigated whether experience with deciding about health or money improves (the quality in terms of consistency of) decision-making. Existing work studying the decisions of physicians (Brosig-Koch et al., 2016) or stockbrokers (Abdellaoui et al., 2013a) found them to be comparable to the decisions typically observed in medicine or economics student samples, respectively. Hence, in Chapter 5, we answered research question 4 by investigating the consistency of decisions about health and money on behalf of others in a student sample consisting of both economics and medical students. We found preference reversals to occur frequently, especially in the health domain. Furthermore, medical students reversed their preferences more frequently overall. Nonetheless, this chapter yielded two strategies to improve consistency. First, although overall preference reversals occurred more frequently for health outcomes, this increase was stronger for economics students than for medicine students. This suggests that domain-relevant experience might improve consistency. Second, valuations for gambles were obtained in two ways: open valuation and by means of completing a guided list of choices. Preference reversals occurred less frequently when guided choice list valuation was used. This suggests that the use of transparent choice-based valuation procedures could increase consistency.

In summary, Part I of this dissertation (Behavioral experiments in health) has shown that:

1. the degree to which individuals are risk averse for small or moderate stakes violates EU,
2. life year losses loom larger than life year gains, and the degree to which this is the case is independent from the quality of life during the years lost or gained,
3. individuals prefer different types and combinations of behaviorally inspired incentives, but these tailored incentives are unrelated to their risk and time preferences,
4. decisions about health and money are both often inconsistent, but preference reversals may be lower in domains in which one has experience and when using choice-based methods.

Applying behavioral insights in health state valuation

Chapters 2 and 3 showed that decisions about health often violate the assumptions underlying EU and the linear QALY model (e.g. Abellan-Perpignan et al., 2006, Attema and Brouwer, 2010a, Bleichrodt and Pinto, 2005, Bleichrodt et al., 1999), which could be problematic as both theories are often (implicitly) assumed to hold in health state valuation with TTO and SG (Brazier et al., 2002, Versteegh et al., 2016). If, regardless of these violations, TTO and SG responses are used assuming EU and the linear QALY hold, the outcomes of these measures are likely to be biased. Bleichrodt (2002), using a theoretical approach, predicted that SG weights are generally too high, while TTO weights are likely to better represent utility of health status (as they are subject to both upward and downward biases). These claims, however, have not yet been substantiated empirically, as determining which method yields more valid QALY weights requires an objective ‘standard’ to compare the utility of health status elicited with both methods to. Given that the degree to which ‘true utilities’ exist is questionable at best (Braga and Starmer, 2005), no decisive argument can be made for any single standard to compare the validity of individuals’ QALY weights to. In Chapter 6 we proposed to test the validity of health state valuations by asking each individual to reflect on their implied QALY weights, given their responses in SG and TTO exercises, themselves. Individuals completed TTO and SG for multiple health states and were subsequently presented with their implied QALY weights. Respondents were asked to indicate for each health state i) whether QALY weights elicited with TTO or SG better reflected the utility of the health state in their opinion, and ii) if the QALY weight should be adjusted upwards or downwards in their opinion if neither method yielded the ‘true’ estimate. The findings of Chapter 6 confirm the predictions made by Bleichrodt (2002) and provide an answer to research question 5: for each health state the majority of individuals indicated that TTO better captured QALY weights. However, on average, QALY weights were adjusted downwards, suggesting some (upward) bias still remains in TTO weights.

In Chapter 7, instead of asking individuals to adjust their QALY weights themselves, another strategy was used. Bleichrodt (2002) proposed that the behavioral insights captured in prospect theory (i.e. loss aversion, utility curvature and probability weighting) may explain the difference between TTO and SG weights. As such, in Chapter 7, we measured these components of prospect theory for each individual and derived TTO and SG weights under prospect theory. We referred to these as ‘corrected weights’, since the bias resulting from violations of EU and the linear QALY model was corrected for, and to the process of adjusting for prospect theory as ‘correction’. When comparing QALY weights, before correction a difference was observed between TTO and SG for all health states, which disappeared after correction. The corrected QALY weights, however, were reduced significantly. As such, the answer to research question 6 is tentatively positive, as it appears that prospect theory can be applied to correct TTO and SG and in our study it yielded the convergence between TTO and SG predicted by Bleichrodt (2002). Furthermore, these results corroborate the suggested conclusions of Chapter 6, i.e. both TTO and SG weights appear to be too high without correction. However, the decrease in QALY weights that resulted from correction for prospect theory was substantial, which compromised the face validity of corrected weights. As such, whether or not correction improved validity of QALY weights was not addressed in this Chapter and remains a topic for future research.

Whereas Chapter 7 showed it is technically possible to correct TTO and SG based on prospect theory, research question 7 (answered in Chapter 8) deals with the feasibility and consequences of using corrected weights in economic evaluation. Using the data from the previous chapter, in Chapter 8 we showed the consequences of using corrected weights instead of uncorrected QALY weights (i.e. classic weights). Obviously, if corrected QALY weights are different from uncorrected ones (i.e. often lower), this may influence estimates of cost-effectiveness and allocation decisions significantly. Several unresolved issues currently preclude the use of corrected weights in practice. First, the influence and validity of the methodology and sample used in Chapter 7 have not yet been tested, which is relevant as other methods have yielded different conclusions (Abdellaoui et al., 2007, Attema et al., 2013). Hence, further replication and methodological exploration is needed, preferably by including respondents from the general public. Second, even though loss aversion and probability weighting are sometimes seen as ‘errors’ or ‘irrationalities’ that need to be corrected for, they could also provide important, policy-relevant information on the emotional value of losses and uncertainties in health. If one believes the latter to be true (as for example argued in: Diecidue and Wakker, 2001, Köbberling and Wakker, 2005), this information might be relevant to incorporate in the context of economic evaluation deliberately. We provided first ideas as to how in Chapter 8. Third, correction based on prospect theory requires assumptions about the reference-point relative to which all outcomes are compared. Whereas in the previous chapters in which prospect theory was applied (i.e. Chapter 2, 4 and 7), this was handled by pragmatically assuming that a single, constant outcome served as reference-point throughout the task, earlier work suggests that reference-points may change from task to task or differ between individuals (van Osch et al., 2006). It has also been argued that individuals may use their subjective life expectancy as reference-points (van Nooten and Brouwer, 2004, van Nooten et al., 2009, Wouters et al., 2015). Hence, before corrected weights can be used in practice, the role and nature of the reference-point should also be explored more extensively.

Chapter 9 commenced this exploration, by studying subjective life expectancy (SLE), i.e. individuals’ expectations with respect to length of life, and in doing so answered research question 8. Earlier work has shown that those with higher SLE (i.e. expecting to become older) generally yield higher QALY weights (Heintz et al., 2013, van Nooten and Brouwer, 2004, van Nooten et al., 2009, van Nooten et al., 2014). Van Nooten et al. (2009) explain this effect by suggesting that SLE serves as reference-point in TTO exercises. Loss aversion may explain higher QALY weights, as for most individuals the durations considered in TTO (e.g. 10 years) are shorter than their SLE. This explanation, however, has not yet been tested and even though SG can also be affected by loss aversion (Bleichrodt, 2002), no work exists testing the effects of SLE on SG. Hence, in Chapter 9 we constructed a model (similar to that of Chapter 7) with SLE as reference-point and apply it to both methods. We derived predictions for TTO and SG through several assumptions about individuals’ preferences (e.g. they are loss averse). These predictions were tested in an experiment, in which individuals’ SLE was used to construct TTO and SG exercises with durations above (i.e. gain versions) and below (i.e. loss versions) their reference-point. Our model predicted that for loss versions TTO weights are subject to upward bias and SG weights are subject to both downward and upward bias (compared to gain versions). These predictions were confirmed in two experiments, in which a sample of students and a sample of individuals aged 60 years and older completed gain and loss versions of TTO and SG. We confirmed the findings of earlier

work, which suggested (based on correlational evidence) that SLE affects TTO. When individuals completed TTO for durations below SLE, they gave up fewer life years compared to TTO with durations above SLE. For SG, these effects were less pronounced, i.e. the effects were smaller compared to TTO (and not significant in the study with an older sample). The results of Chapter 9 suggest that SLE can serve as a reference-point in health state valuation exercises. Hence, if, as we did in Chapter 7, we assume that for all individuals the time in impaired health serves as reference-point, this assumption might not hold for all TTO and SG tasks. Hence, applying a corrective approach based on this assumption might decrease rather than increase the validity of QALY weights.

In Chapter 10, we explored an alternative approach to improve decisions about health. Earlier work in economics has shown that individuals deciding together (e.g. about financial gambles) are less prone to deviate from the predictions of EU (Abdellaoui et al., 2013b), and discount the future less (Denant-Boemont et al., 2017). These effects are hypothesized to result from deliberation, bargaining and information exchange. If such findings can be extrapolated to TTO and SG, the differences between these measures might be reduced by having individuals decide collectively in these tasks (as these differences are partly driven by violations of EU and discounting of future life years). Hence, we developed an experiment aimed at investigating the causal effect of deciding collectively (in dyads) in TTO and SG. We advanced earlier work on this topic (Karimi et al., 2019, Krabbe et al., 1996) by controlling for learning effects, as it has been shown that these affect elicited QALY weights (Augustad et al., 2012). Our work indeed demonstrated the importance of controlling for learning, as almost all effects reported in Chapter 10 were explained by learning, rather than by collective decision-making. In fact, decision quality, QALY weights, and the difference between TTO and SG were unaffected by collective decision-making. As such, although the unresolved issues mentioned in Chapter 8 should be addressed, directly correcting for bias in TTO and SG (as in Chapter 7) appears to be a more promising strategy than collective decision-making.

In summary, Part II of this dissertation has shown that:

1. Although both methods appear to yield QALY weights that are too high, TTO appears to yield more valid QALY weights compared to SG, according to individuals themselves.
2. It is possible to correct TTO and SG weights for prospect theory, and their initially different QALY weights converged after correcting for loss aversion, probability weighting and utility curvature.
3. Although correcting for prospect theory appears to be a promising way forward in health state valuation, several methodological and theoretical challenges currently preclude the use of 'corrected' QALY weights in practice.
4. It appears that individuals' expectations about length of life can serve as reference-point for both TTO and SG.
5. QALY weights are unaffected by deciding collectively, but bargaining and information exchange during collective decision-making yield no benefits beyond learning effects.

Policy implications

The answers to the research questions of this dissertation have several policy implications. First and foremost, the results presented in this dissertation imply that assuming everyone decides rationally about health misrepresents actual decision-making. As such, policy aimed at improving decisions with health consequences can learn from behavioral economics (and the findings reported in this dissertation). In particular, reference-points matter, loss aversion is relatively stable and extends to decisions about health, and people overreact to small probabilities while underreacting to large probabilities (i.e. probability weighting). Indeed, such insights have already reached policy makers, as the past decade has seen a large increase in attention for behavioral public policy (Bernheim and Rangel, 2005, Oliver, 2013a, Oliver, 2015, Shafir, 2013), i.e. public policy based on research in behavioral economics and psychology, rather than being based on traditional economic insights (assuming rationality). Several new public policy tools have been developed, which aim to use behavioral insights to steer people in the ‘right direction’ (e.g. nudges). Such behavioral insights have also been used to ‘supercharge’ existing policy tools, e.g. behaviorally inspired information campaigns or financial incentives (Galizzi, 2014). The findings of this thesis, however, suggest that policy makers should recognize the inherent heterogeneity in how individuals decide about health, both within and between individuals. For a single individual, decisions about health may differ depending on whether risks or delays are involved, and health preferences could depend on how they are elicited. The findings of this dissertation furthermore suggest that large differences exist in, for example, loss aversion or probability weighting between individuals. Hence, general behavioral public policy interventions that aim to benefit from these behavioral insights, e.g. nudges, or information campaigns and financial incentives that are ‘behaviorally inspired’ (Galizzi, 2014) may have heterogeneous effects. Although the (null) results of Chapter 4 suggest that much work remains needed in this area, this dissertation suggests that ‘tailoring’ interventions to this heterogeneity could be a promising way forward (see for example the work on personalized nudges, by: Peer et al., 2019). Such personalization could potentially improve the (cost-)effectiveness of behavioral public policy, by attempting to fit policy to individuals’ heterogeneous preferences.

Part II of this dissertation aimed to apply behavioral insights in a specific context, i.e. health state valuations. Hence, the results of these chapters have several policy implications in this context. First, the results of Chapter 6 suggest that QALY weights based on TTO better reflect individuals’ preferences (according to themselves) than those based on SG. Hence, by extension it could be argued that generic health measures that have health utility tariffs derived with TTO probably more adequately reflect the general public’s preferences for health states than those derived with SG (*ceteris paribus*). As such, this chapter provides some additional arguments for the recommendation of NICE (2018) and the Dutch Health Care Institute (2015) to use EQ-5D (which is valued by TTO) to measure and value quality of life, as opposed to SF-6D (which is valued by SG). However, many individuals, when given the opportunity, still adjusted the implied TTO weights downwards when asked to directly indicate health state utility. As such, reimbursement and allocation decisions using TTO weights (e.g. using nationally representative EQ-5D tariffs: Kim et al., 2016, Versteegh et al., 2016, Xie et al., 2016) may be systematically biased upwards due to the use of QALYs based on TTO (note that Chapters 7 and 9 also suggested that TTO and SG weights are too high). Whether the use of discrete choice experiments in health state valuation (Norman et al., 2013,

Stolk et al., 2010, Xie et al., 2014) offers a way forward to improve validity of QALY weights is, also due to the lack of consensus on how to anchor this method's subjective scale onto the QALY scale (Norman et al., 2016), still open for debate. Hence, in the final chapters of this dissertation alternative ways forward were explored. Chapter 7 suggests that directly correcting biases in TTO and SG may be a promising strategy, but several additional research questions should first be answered before such a corrective approach can be used in policy (see Chapter 8). Collective decision-making (Chapter 10) in health state valuation appears to yield little benefit above and beyond learning effects (which could also be realized through sufficient practice questions, as is for example prescribed in EuroQol Valuation Technology, Stolk et al., 2019).

Limitations and implications for future research

With this thesis I aimed to i) provide additional understanding into how individuals actually decide about health (using theories and methods from behavioral economics), and ii) use this understanding to improve the methods used in health state valuation. My dissertation only provides a partial answer to the research questions that followed from the main aims of my thesis. This is also due to the limitations present in the approaches I used throughout my research. Some important limitations are highlighted below, as well as the potential avenues for future research implied by these limitations.

First, almost all experiments reported in this dissertation utilized a locally recruited student sample, as is usual in economic experiments (e.g. Harrison and List (2004) refer to student samples as 'standard subjects'). Obviously, students differ from the general (or global) population in several ways, i.e., age, education level, current income, residential area (most students are from Rotterdam), and culture. A crucial limitation to take into account when interpreting the findings from this dissertation is, therefore, that the use of a student sample may hamper generalizability of our findings (Henrich et al., 2010). Indeed, the results reported in the Chapters of my dissertation suggest that these student subjects may decide differently about health compared to the (Dutch) general population. For example, Chapters 6, 7, 9 and 10 reported QALY weights that are considerably lower than those observed in a sample representative of the general public. Furthermore, in Chapter 10 this external validity could be compared directly, by using the same methods in both a student sample and sample of people aged 60 years and older. Considerable differences were observed. An important point to make, however, is that most studies reported in this dissertation aimed to compare decisions about health in multiple situations or using different methods (e.g., for TTO or SG for gains and losses, or in groups or individually). Hence, although it may be fair to question whether the *outcomes* of these decisions about health (i.e. the QALY weights or risk attitudes) generalize to other groups than students, raising similar doubts about the decision *processes* implies that students decide about health in a fundamentally different way than other strata in society. For example, all chapters comparing TTO and SG find that although elicited QALY weights were generally lower for students than those observed for the general public (i.e. decision outcomes are not externally valid), as observed in many previous studies (Bleichrodt and Johannesson, 1997, Read et al., 1984, Sackett and Torrance, 1978), TTO weights were significantly lower than SG weights (i.e. externally valid decision process) in both groups.

Second, the work reported in this dissertation almost exclusively relied on lab experiments, i.e. studies in which respondents complete abstract choice tasks in the highly controlled environment of the behavioral lab (the online experiments used in Chapter 4 and 10 are exceptions). Lab experiments have many desirable qualities, i.e. they are affordable means of data collection, allow researchers to maintain high internal validity and as such may be helpful in determining the causal pathways underlying certain effects (Harrison and List, 2004). Lab experiments are often used in behavioral economics and given that the aim of this dissertation was to extend some of the insights from this field to decisions about health, the use of lab experiments can be considered a pragmatic first step. However, decisions about health are often different in many respects from the artificial tasks used in my lab experiments (Galizzi and Wiesen, 2018), and will often take place under less ‘sterile’ conditions. Hence, the use of field experiments, i.e. behavioral research that takes place in the context in which decisions (about health) are actually made (Harrison and List, 2004), appears to be relevant. The literature on risk, time and altruistic preferences for health contains several excellent examples that illustrate how the type of studies reported in Part I of this dissertation could be performed in field contexts (e.g. Brosig-Koch et al., 2016, Galizzi et al., 2016c). It is, however, questionable if the work reported in Part II of this dissertation (i.e. on health state valuation) lends itself for field experiments, as TTO and SG are highly abstract (and unrealistic) decisions about health by design. Extending the findings from Part II to non-student samples, however, may require the use of lab-in-the-field procedures, e.g. mobile labs as in many of the large-scale health state valuation studies used to derive value sets for EQ-5D measures (e.g. Kim et al., 2016, Pickard et al., 2019, Versteegh et al., 2016, Xie et al., 2016).

Third, several of the Chapters of this dissertation reported the results of preference elicitations that may not be feasible for use in other samples or non-lab settings. For example, the non-parametric methodology used in Chapters 3, 5 and 7 has several theoretical advantages over other methods (Abdellaoui et al., 2016, Abdellaoui et al., 2007), but involves a relatively large amount of choices for elicitation of the relevant parameters. As such, the incorporation of this methodology into, for example, large-scale surveys will likely be infeasible, as opposed to survey measures of risk-attitude (e.g. Blais and Weber, 2006) or choice-list methodology (e.g. Holt and Laury, 2002). Furthermore, the methods used in this dissertation require numerate and literate respondents and will not necessarily be applicable in all settings in which the measurement of risk and time preferences is of interest. To accommodate future work in resource-constrained settings, the use of simpler and more efficient elicitation approaches should be explored (i.e. using fewer choices and accessible stimuli, as in: van Wilgenburg, 2018). It is then also crucial to study the extent to which this reduction in complexity increases or decreases the external validity of the measures in predicting decisions about health (e.g. Massin et al., 2018).

Fourth, many of the Chapters in this dissertation have attempted to drop some of the strict assumptions present in EU and the linear QALY model, and instead assume prospect theory to understand decisions about health. Although prospect theory, more than 40 years after its introduction (Kahneman and Tversky, 1979), often is regarded as the best descriptive theory of decision under risk and uncertainty (Wakker, 2010), it is not the only available framework (Kőszegi and Rabin, 2006, Loomes and Sugden, 1982, Wang and Johnson, 2012). As such, many of the findings in this dissertation should be interpreted in light of the limitations of prospect theory. For example, several empirical violations of prospect theory exist, both for

monetary outcomes (Bateman et al., 2007b, Birnbaum, 2006, Payne, 2005) and health outcomes (Feeny and Eng, 2005). Furthermore, prospect theory uses an algebraic approach to describe and predict the *outcomes* of decisions about health, but offers little insight into the *processes* that underly these decisions (Pachur et al., 2017). Some authors have investigated these processes, and their findings suggest that loss aversion and probability weighting may actually reflect how individuals divide their attention among different possible outcomes in decision tasks (Pachur et al., 2014, Suter et al., 2016, Yechiam and Hochman, 2013). The reliance on stated preferences and an algebraic approach to model decisions also typically exclude qualitative methods, although some authors have supplemented their work on prospect theory with think-out-loud procedures (van Osch et al., 2006). Hence, future research could focus more on the processes underlying the phenomena modeled in prospect theory, especially in the context of health state valuation (in which the use of prospect theory appears promising, see Chapter 8).

Concluding remarks

Societies worldwide face increasing proportions of individuals living unhealthily and growing pressure on health care budgets. Throughout this dissertation, I have argued that health economic research can address these worrying trends better by moving beyond *homo economicus*. This traditional model of decisions about health, however desirable one might believe such rationality to be, is unlikely to reflect how real individuals decide. As such, the main goal of this dissertation was to extend and apply methods and theories from behavioral economics to improve understanding of individual and societal decisions about health. Indeed, it appears that many of the insights from behavioral economics extend to decisions about health. Individuals deciding about health (both their own and others') are often inconsistent, are influenced by reference-points, and place excess weight on health losses and overreact to small probabilities while underreacting to large probabilities. However, my hope is that, if the reader indeed derived some new insights about decision about health, it has not made her pessimistic about human decision-making faculties. The inability of individuals in my experiments to live up to the standards of *homo economicus*, in my opinion, if anything indicates the need for additional theoretical and empirical research to better model and understand and influence their decisions.

Summary

Understanding how individual and societal *decisions about health* are made is of crucial importance, given the increasing prevalence of preventable diseases and growing pressure on public health care budgets. Traditionally, in economic models of health-related decision-making it is assumed that individuals decide rationally: i.e. maximizing utility through satisfying consistent and context-independent preferences. Decisions about uncertain health outcomes were modelled with expected utility (EU) theory, which assumes individuals maximize utility while weighting all possible outcomes by their likelihood of occurring. Earlier experimental work has shown that EU is often violated in the context of financial decision-making. In **Part I** of this dissertation several of such behavioral experiments were extended to the health domain, to study if the same behavioral insights apply for decisions about health. In the context of societal decisions about health, EU underlies the measurements required for deriving the preferred outcome in economic evaluations comparing the costs and benefits of medical treatments: Quality-Adjusted Life-Years (QALYs). This measure of health utility is obtained by multiplying the duration of a health gain by a weight that represents the health-related quality of life improvement experienced. QALY weights typically depend on the health state valuation method used for measuring these QALY weights. Two prominent ones are the Time Trade-Off (TTO) method and the Standard Gamble (SG). In **Part II** of this dissertation it is explored if behavioral insights can explain differences between the QALY weights elicited with TTO and SG.

Part I: Behavioral experiments in health

EU is often used to model risk aversion, i.e. to explain why individuals would rather receive nothing than play a gamble with positive expected value, e.g. winning 11 euro with 50% chance or losing 10 euro otherwise. However, one of the famous paradoxes challenging the validity of EU shows that turning down such gambles yields the absurd prediction that gambles with extremely high expected values should also be turned down (e.g. winning 100.000 euro with 50% chance and losing 100 euro otherwise). In **Chapter 2**, this paradox was extended to health by constructing such gambles for health outcomes (e.g. with life duration or disease cases). Our experiment showed that many respondents have risk preferences that violate EU for health. These results suggested that individuals are risk averse for health outcomes, but this risk aversion is not adequately described by EU.

Hence, alternative utility theories should be used to model risk preferences, such as prospect theory. Prospect theory assumes that all outcomes are evaluated compared to a reference-point. Earlier work showed that many individuals are loss averse, i.e. losses (i.e. outcomes below the reference-point) loom larger than gains of the same size (i.e. outcomes above the reference-point). In the experiment reported in **Chapter 3**, individuals' loss aversion was measured for lifetime, where the quality of life during this lifetime was systematically varied. The results reported in this chapter showed that although the majority of respondents were loss averse regardless of quality of life, the strength of this tendency differed strongly between individuals and measurements.

Such heterogeneity in preferences is often observed in behavioral experiments but rarely utilized in policy inspired by such experimental findings. For example, it is unclear how heterogeneity in loss aversion affects the selection or effectiveness of financial incentives inspired by this phenomenon (e.g. deposit contracts). As such, **Chapter 4** explored whether

incentives for physical activity could be tailored, by allowing individuals to design their own financial incentives. The results of this experiment showed that individuals prefer different types and combinations of incentives, but the type of incentives they designed was not associated with risk and time preferences. Therefore, Chapter 4 concluded that it is currently unclear if tailoring incentives to individuals' risk and time preferences is beneficial.

Finally, we explored the consistency of decisions about health (and money) in **Chapter 5**. In this behavioral experiment medicine and economics students' preferences were elicited for health and monetary outcomes using different elicitation methods. The results indicated that preference reversals were more likely to occur for medical students (compared to economics students), and when deciding about health (rather than about money). Furthermore, the results of Chapter 5 showed that the degree of preference reversals depended on the method used for eliciting preferences and on the interaction between field of study and outcome domain, suggesting that experience is associated with higher consistency.

Part II: Applications of behavioral insights to health state valuation

Earlier work suggested that although both TTO and SG can be biased by violations of EU, TTO yields more accurate QALY weights because its outcomes are subject to both upward and downward biases that may cancel out. In **Chapter 6**, we tried to verify this claim by showing individuals their QALY weights measured with TTO and SG and asking them to reflect on these weights. According to individuals themselves, TTO yielded more valid QALY weights than SG. However, both methods yielded QALY weights that were too high according to respondents, as they were on average adjusted downwards. As such, deriving QALY weights with TTO or SG could overestimate the utility assigned to health states, which could have a distorting influence on economic evaluations that rely on these weights.

In **Chapter 7** we explored if prospect theory could be used to 'correct' these biases. In an experiment, next to completing TTO and SG tasks, each respondent also completed measurements of loss aversion, probability weighting (i.e. if a respondent overestimates or underestimates probabilities) and the shape of their utility function. If we took into account these individual differences captured by prospect theory, the differences between TTO and SG disappeared for all health states. Although this convergence of TTO and SG is promising and in accordance with earlier theoretical predictions, QALY weights for both methods were significantly reduced. This suggests the need for further validation of the correction process.

Next, **Chapter 8** explored the challenges currently precluding the use of such a 'corrective approach' in economic evaluation. Seeing as the use of corrected weights could strongly affect economic evaluation, the results reported in Chapter 7 should first be replicated with different samples and methodologies. Furthermore, although probability weighting and loss aversion for health signal that uncertainties and health losses are of significant importance, whether and how this should be incorporated in economic evaluation is unclear. Finally, since applying a corrective approach requires assumptions about the reference-point in prospect theory, the role and nature of reference-points in health should also be further explored.

Chapter 9 commenced this exploration of the role of the reference-point by investigating the influence of subjective life expectancy (SLE). In earlier work using TTO it was observed that those with higher SLE gave up fewer life years, which was posited to result from loss aversion. In this chapter we provided the theoretical foundation to experimentally test this prediction for both TTO and SG. Respondents' SLEs were used to construct TTO and SG

tasks with life years above or below SLE. As predicted, subjects gave up fewer years in TTO and were less risk-tolerant in SG below SLE, suggesting that years below SLE are seen as losses and SLE can serve as reference-point. The short durations typically used in TTO and SG could, therefore, yield upward bias and it is unclear how this should be addressed.

Finally, we explored an alternative solution to biases in TTO and SG in **Chapter 10**: deciding collectively. Earlier work suggested that collective decision-making could reduce violations of EU, meaning that the difference between weights derived with TTO and SG resulting from such violations might also diminish. Nonetheless, the findings of the experiment reported in this chapter suggested that collective decision making in dyads has little effect on decision quality. Furthermore, QALY weights remained similar between individual and collective decisions and the typical difference in elicited weights between TTO and SG was not affected.

Collectively, Part I and Part II have several policy implications. First, as many experimental findings could be extended to the health domain, the increased interest in policy interventions including behavioral insights is supported by this dissertation. Our findings, however, suggest policy makers should recognize that large differences exist between individuals in terms of how they decide about health, and as such, personalized interventions tailored to individuals' heterogeneous preferences may be warranted. Second, this dissertation showed that TTO and SG are likely to lead to QALY weights that are biased upwards. The research reported in this dissertation suggested that applying a corrective approach could be a promising way forward, although several additional research questions should first be answered. Such future research efforts should aim to avoid some of the limitations present in most of the chapters of this dissertation. That is, future work could explore the validity of the conclusions reported in this dissertation with general public samples, in (lab in the) field settings, with more efficient measurement methods, and using alternative theories or qualitative methods.

Concluding, this dissertation showed that individuals deciding about health are often inconsistent, are influenced by reference-points, and place excessive weight on health losses and extreme probabilities. Currently, the biases these preferences may yield are disregarded, as they violate traditional assumptions about rationality or for pragmatic reasons. I believe that additional theoretical and empirical research is needed to better model these decisions and hope my dissertation has contributed to a better understanding of decisions about health.

Nederlandse samenvatting

De groei in prevalentie van te voorkomen ziektes en de druk op publieke zorgbudgetten maken het van cruciaal belang te begrijpen hoe beslissingen over gezondheid zowel door individuen als maatschappijen worden genomen. Wanneer in de economische traditie modellen worden gebruikt om aan gezondheid gerelateerd gedrag te voorspellen, werd aangenomen dat iedereen rationeel beslist. Dat wil zeggen dat eenieder nut of *utility* maximaliseert door aan consistente en context-onafhankelijke voorkeuren te voldoen. Onzekere beslissingen over gezondheid werden gemodelleerd aan de hand van de verwachte nutstheorie, dat wil zeggen *expected utility (EU) theory*, waarin wordt aangenomen dat individuen alle uitkomsten wegen op basis van de kans dat ze voorkomen. Eerder experimenteel onderzoek heeft uitgewezen dat de aannames gebruikt in EU vaak worden geschonden wanneer men keuzes over geld maakt. In **Deel I** van dit proefschrift werden enkele van deze bevindingen vertaald naar het gezondheidsdomein, om te onderzoeken of dezelfde gedragsinzichten van toepassing zijn op beslissingen over gezondheid. EU ligt ook ten grondslag aan maatschappelijke beslissingen over gezondheid. Er wordt namelijk aangenomen dat EU standhoudt bij het meten van de uitkomstmaat die in economische evaluaties van medische behandelingen bij voorkeur wordt gebruikt: voor kwaliteit gecorrigeerde levensjaren (Quality-Adjusted Life Years, QALY's vanaf nu). Deze uitkomstmaat wordt geschat door de gezondheidswinst van medische behandelingen uit te drukken als het product van de lengte en kwaliteit van die levenswinst. De voor die berekening benodigde kwaliteitsgewichten, zo blijkt uit eerder werk, zijn afhankelijk van de methode waarmee ze worden geschat. Time trade-off (TTO) en standard gamble (SG) zijn populaire methodes voor dit doel. In **Deel II** van dit proefschrift werd onderzocht of de tussen TTO en SG verschillende kwaliteitsgewichten aan de hand van gedragsinzichten kunnen worden verklaard.

Deel I: Gedragsexperimenten voor gezondheid

EU wordt vaak gebruikt om risicoaversie te modelleren, dat wil zeggen om uit te leggen waarom men liever niets ontvangt dan een gok neemt met positieve verwachte waarde, e.g. het winnen van 11 euro met 50% kans of anders 10 euro verliezen. Echter, een van de meest bekende paradoxen die de validiteit van EU in twijfel trekt, laat zien dat het afwijzen van zulke gokken tot de absurde voorspelling leidt dat gokken met extreem hoge verwachte waarde ook worden afgewezen (e.g. 100.000 euro met 50% kans of anders 100 euro verliezen). In **Hoofdstuk 2** werd deze paradox vertaald naar het gezondheidsdomein door het aanbieden van zulk soort gokken met gezondheidsuitkomsten (e.g. met levensduur of ziektegevallen). Ons experiment liet zien dat men vaak risicovoorkeuren heeft die niet in overeenstemming zijn met EU. Deze resultaten suggereren dat men risicoavers is wat betreft gezondheidskomsten, maar dat deze risicoaversie door EU niet adequaat wordt beschreven.

Zodoende lijkt het nodig om andere nutstheorieën te gebruiken om risicovoorkeur te modelleren, zoals *prospect* theorie. In prospect theorie wordt aangenomen dat uitkomsten vergeleken worden met een referentiepunt, waar alles hierboven een winst en daaronder een verlies is. Eerder werk toonde aan dat veel mensen verliesafkerig zijn: dat wil zeggen een verlies weegt zwaarder dan evenredige winst. In het experiment dat in **Hoofdstuk 3** werd besproken werd verliesafkeer voor levensduur gemeten, waar de levenskwaliteit waarin die levensduur werd doorgebracht gevarieerd werd. De resultaten van dit hoofdstuk lieten zien

dat het merendeel van de respondenten verliesafkerig was, ongeacht de levenskwaliteit van de jaren waarmee gemeten werd. Er waren wel aanzienlijke verschillen tussen individuen en metingen.

Dit soort heterogeniteit wordt in de gedragsexperimenten vaak waargenomen, maar wordt zelden in de praktijk ingezet voor beleid dat door experimentele bevindingen is geïnspireerd. Het is bijvoorbeeld onduidelijk of de verschillen in verliesafkeer invloed hebben op de effectiviteit van financiële prikkels die met dit inzicht in gedachten zijn ontworpen (bijv. *deposit* contracten). Zodoende werd in **Hoofdstuk 4** verkend of financiële prikkels op maat gemaakt kunnen worden, door deelnemers te vragen hun eigen financiële prikkels te ontwerpen. Dit experiment liet zien dat er grote verschillen bestaan in het soort prikkels dat men zou willen ontvangen, maar dat die verschillen niet te verklaren zijn aan de hand van tijds- en risicovoorkeuren. Zodoende werd op basis van Hoofdstuk 4 niet duidelijk of het op maat maken van financiële prikkels direct een toegevoegde waarde heeft.

Ten slotte werd in **Hoofdstuk 5** de consistentie van keuzes over geld en gezondheid onderzocht, dat wil zeggen *preference reversals*. Voor dit onderzoek werd een steekproef van zowel economie- als geneeskundestudenten geworven, bij wie met verschillende methoden voorkeuren werden gemeten voor uitkomsten in zowel het financiële alsmede het gezondheidsdomein. De resultaten van dit hoofdstuk toonden dat een *preference reversal* vaker voorkwam onder geneeskundestudenten (ten opzichte van economiestudenten) en voor keuzes over gezondheid (ten opzichte van keuzes over geld). Daarnaast wezen de resultaten van Hoofdstuk 5 uit dat de mate van consistentie afhangt van de methode die wordt gebruikt om voorkeuren te meten en de interactie tussen studierichting en uitkomstdomein. Dit laatste suggereerde dat hogere consistentie met ervaring samenhangt.

Deel II: de toepassing van gedragsinzichten voor gezondheidswaardering

Dit hoofdstuk bouwde voort op eerder werk dat stelde dat, ondanks dat zowel TTO als SG door afwijkingen van EU vertekend kunnen zijn, TTO een betere weergave zou geven van de waarde van gezondheidstoestanden. De verklaring die werd voorgesteld is dat vertekeningen bij TTO in tegengestelde richting (zowel opwaarts als neerwaarts) werken en elkaar misschien opheffen. In **Hoofdstuk 6** poogden wij deze stelling te testen door een experiment uit te voeren waarin respondenten hun kwaliteitsgewichten, gemeten met TTO en SG, te zien kregen en op deze gewichten reflecteerden. Kwaliteitsgewichten gemeten met TTO leken meer valide te zijn volgens respondenten zelf. Echter, beide methoden leverden te hoge kwaliteitsgewichten op volgens respondenten zelf; ze werden namelijk naar beneden bijgesteld. Zodoende suggereerden deze resultaten dat het meten van kwaliteitsgewichten met TTO en SG gemiddeld te hoge resultaten oplevert, wat een vertekende invloed kan hebben op economische evaluaties (waarin deze gewichten worden toegepast).

In **Hoofdstuk 7** werd daarom onderzocht of prospect theorie gebruikt kan worden om deze vertekening te ‘corrigeren’. In dit experiment werden TTO en SG taken afgenomen, en werden verschillende componenten van prospect theorie gemeten: verliesafkeer, kansweging (d.w.z. de mate waarin men kansen onder- of overschat) en de nutsfunctie voor levensduur. Als deze individuele kenmerken in de schatting van kwaliteitsgewichten op basis van TTO en SG werden meegenomen, verdween het problematische verschil tussen de uitkomsten van beide methodes. Alhoewel deze convergentie veelbelovend lijkt en overeenkomt met de vooropgestelde verwachtingen, waren de resulterende gecorrigeerde kwaliteitsgewichten

aanzienlijk lager. Deze reductie laat de noodzaak zien om het correctieproces verder te valideren.

Vervolgens bezigde **Hoofdstuk 8** zich met de uitdagingen die op dit moment het gebruik van deze ‘correctieve aanpak’ in de praktijk uitsluiten. Omdat het gebruik van gecorrigeerde gewichten economische evaluaties sterk zou kunnen beïnvloeden, moeten de bevindingen van Hoofdstuk 7 eerst gerepliceerd worden met andere steekproeven en meetmethoden. Daarnaast is, ondanks dat kansweging en verliesafkeer uitwijzen dat risico’s en verliezen op gezondheidsgebied een belangrijke rol spelen, het niet duidelijk hoe deze inzichten in economische evaluaties mee te nemen. Bovendien, aangezien een correctieve aanpak aannames over het referentiepunt in prospect theorie nodig maakt, moet meer onderzoek gedaan worden naar de rol en het ontstaan van referentiepunten op het gebied van gezondheid.

In **Hoofdstuk 9** werd met onderzoek naar de rol van het referentiepunt van start gegaan door de invloed van de subjectieve levensverwachting verder uit te diepen. In eerder werk waar TTO taken werden afgenomen, werd geobserveerd dat wanneer men langer verwachtte te leven, er minder jaren opgegeven werden. Dit werd door verliesafkeer verklaard. In dit hoofdstuk verschaften we de theoretische onderbouwing om deze verklaring te testen in een experiment, zowel voor TTO als SG. De subjectieve levensverwachting van deelnemers werd gebruikt om TTO en SG taken op te stellen met levensjaren volledig onder of boven die levensverwachting. Zoals voorspeld werden er in TTO voor jaren onder de subjectieve levensverwachting minder jaren opgegeven en werd in SG minder risico genomen. Deze bevindingen suggereerden dat jaren onder de subjectieve levensverwachting als verliezen worden gezien, en de subjectieve levensverwachting dus een referentiepunt was voor zowel TTO als SG. De korte levensduren die normaliter in beide methodes worden gebruikt, zouden zodoende een verhogend effect op kwaliteitsgewichten kunnen hebben.

Ten slotte werd in **Hoofdstuk 10** een alternatieve oplossing voor het verkleinen van vertekening in TTO en SG voorgesteld: collectieve besluitvorming. Eerder werk liet zien dat gezamenlijk beslissen tot een verminderde afwijking van EU kan leiden, wat zou kunnen betekenen dat het door de afwijkingen van EU verklaarde verschil tussen TTO en SG ook zou kunnen afnemen. Echter, de bevindingen van een experiment lieten geen effect zien van collectieve besluitvorming op de kwaliteit en uitkomsten van beslissingen in TTO en SG. Ook het verschil tussen de uitkomsten van TTO en SG bleef onveranderd.

Deel I en Deel II hebben enkele gezamenlijke beleidsimplicaties. Ten eerste, aangezien veel van de experimentele bevindingen zich lieten vertalen naar keuzes over gezondheid, ondersteunt dit proefschrift de vergrote interesse in beleidsinterventies waarin gedragsinzichten worden meegenomen. Onze bevindingen suggereren daarentegen wel dat beleidsmakers zullen moeten aannemen dat er grote individuele verschillen bestaan in de manier waarop over gezondheid wordt besloten. Zodoende lijkt het gebruik van gepersonaliseerde interventies, die op basis van individuele voorkeuren op maat gemaakt worden, gerechtvaardigd. Ten tweede laat dit proefschrift zien dat TTO en SG in de gebruikelijke vorm waarschijnlijk vertekende resultaten opleveren. Ons onderzoek suggereert dat een aanpak op basis van individuele correctie veelbelovend is, maar dat enkele openstaande vragen eerst verder onderzocht moeten worden. Dat wil zeggen, toekomstig onderzoek zou zich kunnen richten op het verkennen van de validiteit van onze bevindingen

in representatieve steekproeven, buiten het experimentele lab, met efficiëntere meetmethodieken en door alternatieve theorieën of kwalitatieve methoden te gebruiken.

In conclusie, mijn onderzoek liet zien dat wanneer men beslist over gezondheid, inconsistenties vaak voorkomen, referentiepunten invloed hebben en kleine verschillen in kansen en verliezen aanzienlijk gewicht hebben. De vertekening die deze bevindingen tot gevolg hebben, wordt op dit moment vaak genegeerd omdat deze niet aansluit bij aannames over rationaliteit, of uit pragmatisch oogpunt. Ik geloof dat meer theoretisch en empirisch onderzoek nodig is om deze beslissingen over gezondheid beter te modelleren en te begrijpen en ik hoop dat mijn proefschrift daar een bijdrage aan heeft geleverd.

Portfolio

This portfolio contains an overview of activities and accomplishments during my Ph.D.

Training

2019	Principal investigator training for health state valuation (EuroQol group)
2019	Self-presentation: Focus, structure and visualization (Erasmus Graduate School for Social Sciences and Humanities)
2019	Analytical Storytelling (Erasmus Graduate School for Social Sciences and Humanities)
2019	Storytelling in Education (Risbo teaching institute, Rotterdam)
2019	Group Dynamics (Risbo teaching institute, Rotterdam)
2018	Powerful Presenting: Pecha Kucha format (Erasmus University Rotterdam)
2018	Behavioral Experiments in Health Summer School (University of Cologne)
2018	Bounded Rationality Summer Institute (Max Planck Institute, Berlin)
2018	Basic Didactics (Risbo teaching institute, Rotterdam)
2018	Theatre techniques in education (Risbo teaching institute, Rotterdam)
2018	Making an Academic Poster that Stands Out, EGSB (Erasmus Graduate School of Social Sciences and Humanities)
2018	Risk & Rationality, (Tinbergen Institute, Amsterdam)
2017	Basic Didactics for T.A.'s (Risbo teaching institute, Rotterdam)
2017	Behavioral Insights Summerschool, (Erfurt University)
2017	Advanced Behavioral Economics, Erasmus School of Economics, Rotterdam
2017	PhD workshops: 'Your Presentation and You' & 'Finding the Social Value of Your Research'
2016/2017	Academic Writing in English, Erasmus University Rotterdam
2016/2017	PhD Project Management, Erasmus University Rotterdam

Teaching

2019-2020	Lecturer, working groups and coordination in elective minor Understanding Health Behavior: Insights and applications from behavioral and health economics
2018-2020	Lecturer and working groups (and coordination ad interim) in Behavioral Decision Theory in Health/Advanced Economic Evaluation, elective course at M.Sc. level (Erasmus School of Health, Policy & Management)
2016-2018	Working groups for Methods and Techniques II, Bachelor Health Sciences (Erasmus School of Health, Policy & Management)
2016-2019	Workgroups and supervision Quantitative research skills premaster program (Erasmus School of Health, Policy & Management)
2016-2019	Thesis supervision (at bachelor and master level)

Awards, Grants and Scholarships

2020	Research grant. EuroQol foundation (principal investigator)
2020	Research grant. EuroQol foundation (co-investigator)
2020	Research grant. EuroQol foundation (co-investigator)
2020	Research visit grant. Erasmus Trustfonds.

Portfolio

2020	Academy van der Gaag grant. Royal Dutch Academy for Arts and Sciences
2019	Research grant. EuroQol foundation (principal investigator)
2019	Research grant. EuroQol foundation (co-investigator)
2019	Research grant. EuroQol foundation (principal investigator)
2019	Research grant. EuroQol foundation (co-investigator)
2019	Research grant. EuroQol foundation (principal investigator)
2018	Runner up Poster prize at Bounded Rationality Summer Institute 2018 at Max Planck Institute Berlin (chairs: Gerd Gigerenzer & Ralph Hertwig)
2018	Scholarship for participation at Bounded Rationality Summer Institute 2018 at Max Planck Institute Berlin

Presentations

Invited seminars

Lipman, S.A. & Pachur, T. Attention allocation in the valuation of health: are decision processes in health state valuation in accordance with prospect theory? Centre for Adaptive Rationality, Max Planck Institute Berlin, July 2020.

Lipman, S.A. One size fits all? Designing financial incentives tailored to individual economic preferences. Amsterdam Medical Center, February, 2020

Lipman, S.A. One size fits all? Designing financial incentives tailored to individual economic preferences. Leiden University, September 2019

Lipman, S.A., Brouwer, W.B.F. & Attema, A.E. Living up to expectations - Experimental tests of subjective life expectancy as reference-point in time trade-off and standard gamble. Paper presented at Erasmus School of Economics, department of Applied Economics, May, 2019.

Conference and summer school presentations

2020:

EuroQol Early Career Researcher Meeting, Prague;

Workshop in Behavioral and Experimental Health Economics, Innsbruck;

Association for Researchers in Psychology Conference, Egmond aan Zee.

2019:

Subjective Probability, Utility, and Decision Making Conference, Amsterdam;

International Health Economics Association Biennial World Congress, Basel;

Welfare Improvement through Nudging Knowledge Conference; Utrecht

Lowlands Health Economics Study Group, Almen.

2018:

European Health Economics Association conference in Maastricht;

1st Summer School on Behavioral Experiments in Health, University of Cologne;

Summer Institute on Bounded Rationality, Max Planck Institute, Berlin;

Biennial European Conference on Medical Decision Making, Leiden;

Lowlands Health Economics Study Group, Hoenderloo

2017:

European Health Economics Association PhD Student-Supervisor Conference, Lausanne;

International Health Economics Association Biennial World Congress, Boston;

Behavioral Insights Summer School, Erfurt;

Lowlands Health Economics Study Group, Rotterdam.

List of Publications

In this dissertation:

Lipman, S.A., Brouwer, W.B.F. & Attema, A.E. (2020). What is it going to be, TTO or SG? A direct test of the validity of health state valuation. *Health Economics*

Lipman, S.A. (2020). One size fits all? Designing financial incentives tailored to individual preferences. *Behavioural Public Policy: 1-15*.

Lipman, S.A., Brouwer, W.B.F., & Attema, A.E. (2020). Living up to expectations: Experimental tests of subjective life expectancy as reference point in time trade-off and standard gamble. *Journal of Health Economics, 102318*

Attema, A.E., Bleichrodt, H., l'Haridon, O., & Lipman, S.A. (2020). Comparison of individual and collective decision making for time trade-off and standard gamble. *European Journal of Health Economics, 1-9*.

Lipman, S. A., Brouwer, W. B.F., & Attema, A. E. (2019). QALYs without bias? Nonparametric correction of time trade-off and standard gamble weights based on prospect theory. *Health Economics, 28(7)*, 843-854.

Lipman, S. A., Brouwer, W.B.F., & Attema, A. E. (2019). The corrective approach: Policy implications of recent developments in QALY measurement based on prospect theory. *Value in Health, 22 (7)*, 816-821

Lipman, S.A., Brouwer, W.B.F., & Attema. (2019) - A QALY LOSS IS A QALY LOSS IS A QALY LOSS: Independence of loss aversion from health states. *European Journal of Health Economics, 20 (3)*, 419-426. doi: 10.1007/s10198-018-1008-9

Lipman, S. A., & Attema, A. E. (2019). Rabin's paradox for health outcomes. *Health economics, 28(8)*, 1064-1071

Other academic publications:

Lipman, S. A., & Attema, A. E. (2020). Good things come to those who wait—Decreasing impatience for health gains and losses. *PLoS ONE, 15(3)*, e0229784.

Attema, A.E. & Lipman, S.A. (2018) - Decreasing impatience for health outcomes and its relation with healthy behavior. *Frontiers of Applied Mathematics*. doi: 10.3389/fams.2018.00016

Lipman S.A., Burt S.A. (2017). Self-reported prevalence of pests in Dutch households and the use of the health belief model to explore householders' intentions to engage in pest control. *PLoS ONE 12(12)*: e0190399.

Professional publications:

Bonfrer, I., & Lipman, S.A. (2020). Geef alle zorgprofessionals dezelfde genereuze bonus. *Het Parool*, 18-18.

Lipman, S.A. (2019). Putting a price on life: why and how. Opent extern (blog). United Academics Magazine. (available: 21 June 2019): <https://www.ua-magazine.com/health-care-putting-price-life/>

Lipman, S.A. (2019). What would you do if you have 10 years to live? (blog). Erasmus School of Health Policy & Management website (available: 23 May 2019).: <https://www.eur.nl/en/eshpm/news/value-health-what-would-you-do-if-you-have-10-years-live>

Burt, S.A. & Lipman, S.A. (2018) Dierplagen in Nederlandse woningen. Intenties van bewoners ten opzichte van bestrijding. *Dierplagen Informatie*, 1.

Miscellaneous

Professional activities: Referee for *Value in Health*, *Journal of Health Economics*, *European Journal of Health Economics* & *Health Economics*, *Health Economics Policy & Law*, *PLOS ONE*, organizer of monthly Ph.D. discussion meeting (2016-2019), board member for young-ESHPM (representing Ph.D. students at Erasmus School of Health Policy & Management from 2017-2019), organizer of ESHPMs Ph.D. Platform 2018, local organizing committee for 2nd Behavioral Experiments in Health Summer School (July 2019, Rotterdam)

Teaching certification: University Teaching Qualification (Risbo teaching institute, Rotterdam)

Volunteering: Coach, trainer, and referee coordinator at volleyball club VollenGo in Gouda.

Dankwoord

Maanden voordat ik begon met het schrijven van een introductie van mijn proefschrift ben ik gestart met mijn dankwoord, ik denk het belangrijkste onderdeel van mijn proefschrift. In de fantastische tijd die ik heb doorgemaakt gedurende het promoveren zijn er veel mensen die ik graag wil bedanken. Zij hebben een cruciale rol gehad in de totstandkoming van dit proefschrift, omdat ze het promoveren gemakkelijker maakten, of omdat ze het juist moeilijker maakten.

Te beginnen met hen die het promoveren gemakkelijker maakten:

Allereerst wil ik Arthur en Werner bedanken, mijn (co)promotors, begeleiders en mentoren.

Werner, ik begon als promovendus toen jij nog (officieel) decaan van iBMG was. Je wist me gelijk voor deze faculteit te winnen door jouw enthousiasme voor interdisciplinair onderzoek, maar vooral door de openheid, interesse en humor waarmee je met mij in gesprek ging. In mijn tweede sollicitatieronde kwam ik onverhoopt een half uur te laat. Bezweet en gestresst kwam ik aan, en daar zaten jullie rustig keuvelend muziek te luisteren, niet gestoord door mijn vertraging. Binnen een seconde voelde ik me op mijn gemak bij je, en dat is niet meer weg gegaan. Onze gesprekken waren een maandelijks hoogtepunt, en je was altijd in staat om mij te helpen richting te geven aan de wirwar van plannen en ideeën waarmee ik aankwam – waar je mijn interesse altijd leidend liet zien.

Arthur, meer dan vier jaar geleden nam je de telefoon op, en stimuleerde je een jonge psycholoog om te solliciteren op jouw vacature. Meer dan enthousiasme, de wil om te leren en een totaal andere achtergrond had ik jou op dat moment niet te bieden. Het vertrouwen dat je me hebt gegeven door mij aan te nemen als je eerste promovendus is enorm, en ik hoop dat ik de verwachtingen die je van te voren had waar heb kunnen maken. Ik kon elke dag bij je terecht, voor koffie, voor vragen of advies, of voor je dubieuze kijk op voetbal. Dank voor je analytisch vermogen, je aandacht voor detail, jouw geduld, en voor het bieden van het platform waar ik mijn ambitie kan waarmaken.

Ik ben ook veel dank verschuldigd aan mijn andere coauteurs: Han, Olivier, en Sebastian.

Han, je hebt op afstand een grote rol gespeeld in mijn ontwikkeling. De methoden die je hebt ontwikkeld vormen de basis van veel van mijn werk. Dat ik samen met jou heb kunnen werken aan een paper en een deel van jouw onderwijs in *Advanced Economic Evaluation* over heb mogen nemen voelt als een enorme eer.

Olivier, I was very lucky that you were in Rotterdam so often. Thanks for your patience, especially when working through the analysis of our paper. Your proficiency in data-analysis is especially impressive, and I have learnt so much from just watching your work.

Sebastian, thanks for keeping me caffeinated at all times. We worked on several projects together closely, which is testament of your generosity – you were willing to let me join in on some of your many good ideas.

Verder wil ik graag een aantal mensen bedanken die over de jaren heen bijdrages hebben geleverd aan mijn werk, door het bieden van praktische ondersteuning of het lezen en becommentariëren van eerdere versies van de hoofdstukken in dit proefschrift. Ten eerste,

dank aan Marcel en Christiaan van Erasmus Behavioral Lab, waar ik een groot deel van de experimenten heb uitgevoerd. Mijn dank gaat ook uit naar de beheerders van het Erasmus Research Participation System, waar ik voor het merendeel van mijn hoofdstukken deelnemers heb geworven. Onze secretaresse Liza dank ik voor haar hulp en het oppakken van dingen waar ik geen verstand van heb, zoals het boeken van vluchten, verzekeringen (voor vermiste koffers) en ga zo maar verder. Via deze weg bedank ik ook graag opnieuw de vele discussianten en vrijwilligers die mijn stukken lazen en van feedback voorzagen!

Mijn grote dank gaat ook uit naar de groep Behavioral Economics, in het bijzonder naar Peter Wakker. Elke week mocht ik aansluiten bij jullie wekelijkse bijeenkomst waar een mix tussen nieuwe en klassieke papers werden gelezen en besproken. Ik heb me altijd onderdeel van jullie groep gevoeld, ondanks dat ik dit feitelijk niet ben. De inzichten die ik in aan deze bijeenkomsten heb ontleend zijn cruciaal geweest in mijn verdere vorming.

Ik vervolg mijn dankwoord graag met het adresseren van hen die het promoveren moeilijker maakten, maar wel leuker, gezelliger en uitdagender – zonder jullie had ik het nooit (zo snel) afgemaakt.

Meg, Mariska en Friederike, het grootste deel van mijn promotietraject waren jullie mijn kamergenoten. Dank voor het tolereren van mijn rommelige bureau, slechte humor en hardop uitgesproken warrige gedachtegangen. Meg, thanks for being always being thoughtful, at times snarky and often sarcastic – and allowing me to practice my English all day long. Mariska, dank de gezelligheid! We hebben veel gelachen, onze vlucht uit de stromende regen richting het sportcafé in Boston was op dat gebied een hoogtepunt. Friederike, van een student in mijn eerste hoorcollege werd je mijn collega en kamergenoot. Je goede humeur deed me altijd goed en ik ben blij dat je nu bij RIVM een mooie nieuwe uitdaging kan aangaan!

Ik heb daarnaast het geluk gehad om te mogen werken op de afdeling Health Economics in het midden van een grote groep uitzonderlijke wetenschappers en fijne collega's, die de dagen op kantoor misschien niet productiever maar wel uit te houden maakten. Leander, voor vele borrels heb je me kunnen enthousiasmeren, zelfs tot in een wolkenkrabber in Boston. Jannis, thanks for teaching me how to play squash and for wiping the floor with me week after week until I did the same with you. I am very happy you wanted to be a paranimf at my defence! Pieter B., we werkten niet samen, maar tijdens de pauzes gaf je me een aantal keer carrière advies wat ik dankbaar ter harte heb genomen. Pieter van B., Feyenoord-supporter zijn was de afgelopen jaren vaak geen pretje, en gelukkig was ik niet de enige die op maandag vaak chagrijnig op werk kwam. Vivian, dank voor je zwarte humor, ik ben blij dat het aan het einde van onze promotietrajecten toch nog gelukt is samen een paper te schrijven. Judith, dankzij jou was ik gelukkig niet de enige 'crazy cat person' op de afdeling. Klas, thanks for advice on movies and series, and for stretching out lunch breaks to finish your enormous lunches. Job, dank voor de kansen en vrijheid die je me gaf in het ontwikkelen van de minor. Marianne and Igna, I feel lucky to have such accomplished colleagues present to ask for advice for the next steps of my career.

Ik heb mijn promotietraject doorlopen met een grote groep andere promovendi. Maandelijks spraken we af ons werk in ontwikkeling te bespreken, en via deze bijeenkomsten heb ik de

breedte van mijn vakgebied leren kennen. Thanks to all participants of these HE-junior meetings over the years, for giving insight into your work and providing your insights into mine: Joaquim, Linda, Jenny, Wally, Sara, Charlotte, Sebastian, Pugo, Sebastian, Marlies, Valerie, Samare, Friederike, Rita, Lisa, Jawa and Dilnoza.

Ik ben ook veel dank verschuldigd aan Elly (en de EuroQol groep), door wiens hulp en advies het mij gelukt is om een aantal fondsen te bemachtigen die me in staat stellen blijven werken aan het begrijpen en verbeteren van de methodologie voor het waarderen van gezondheidstoestanden. Ik kijk erg uit naar de mogelijkheid om te kijken of een deel van de inhoud van dit proefschrift ook in de praktijk ingezet kan worden.

Een grote eer was het om de belangen van promovendi te behartigen binnen onze faculteit met young-ESHPM. Timo, Eline, Mathilde en Marthe, wat ben ik trots op wat wij in een korte periode hebben kunnen organiseren en bereiken – en door dit met jullie te doen voelde het nooit als (hard) werken.

Ik kan natuurlijk mijn teamgenoten, vrienden en trainers bij volleybalclub VollenGo de afgelopen jaren niet vergeten. In het bijzonder bedank ik Reinier, Frank, Erik en Gerard voor ‘doen-we-er-nog-eentje?’ en de vele thuiswerkdagen op vrijdag.

Timo, Michael en Maurice, dank voor jullie vriendschap over de jaren heen. Ik hoop dat jullie me nog vele jaren van mijn werk af houden.

Om af te sluiten, mijn eeuwige dank en liefde gaan uit naar mijn familie.

Papa, veel jongetjes willen wel worden zoals hun vader – en ik probeer het nog steeds. Je hebt me nooit aangemoedigd om de wetenschap in te gaan, dat was niet nodig, je bent mijn voorbeeld. Mam, jouw aanmoediging en onvoorwaardelijke trots maken alles wat ik bereikt heb nog mooier – en gaven me het vertrouwen in mezelf dat zo cruciaal was tijdens het promoveren. Rosa, als je me om hulp met statistiek vroeg voelde ik me altijd vereerd, ik ben heel blij dat we ook jouw afstuderen binnenkort kunnen vieren. Niels, veel dank dat je paranimf bij mijn verdediging wil zijn, ik kijk er naar uit om je in rokkostuum te zien.

Lieve Ilse, jou bedanken is het moeilijkste van allemaal. Dank voor het delen van mijn hart en mijn thuis, voor je steun en toeverlaat, voor het zijn van een reden om hard te werken en gauw naar huis te komen. Jij bent mijn referentiepunt: ten opzichte van bij jou zijn is al het andere een verlies.

References

- ABDELLAOUI, M. 2000. Parameter-free elicitation of utility and probability weighting functions. *Management Science*, 46, 1497-1512.
- ABDELLAOUI, M., BLEICHRODT, H. & KAMMOUN, H. 2013a. Do financial professionals behave according to prospect theory? An experimental study. *Theory and Decision*, 74, 411-429.
- ABDELLAOUI, M., BLEICHRODT, H. & L'HARIDON, O. 2008. A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty*, 36, 245.
- 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., DIECIDUE, E. & ÖNCÜLER, A. 2011. Risk preferences at different time periods: An experimental investigation. *Management Science*, 57, 975-987.
- ABDELLAOUI, M., L'HARIDON, O. & PARASCHIV, C. 2013b. Individual vs. couple behavior: an experimental investigation of risk preferences. *Theory and Decision*, 75, 175-191.
- ABELLAN-PERPIÑAN, J. M., BLEICHRODT, H. & PINTO-PRADES, J. L. 2009. The predictive validity of prospect theory versus expected utility in health utility measurement. *Journal of Health Economics*, 28, 1039-1047.
- ABELLAN-PERPIÑAN, J. M., PINTO-PRADES, J. L., MENDEZ-MARTINEZ, I. & BADIAL-LACH, X. 2006. Towards a better QALY model. *Health Economics*, 15, 665-76.
- ADAMS, J., GILES, E. L., MCCOLL, E. & SNIHOTTA, F. F. 2014. Carrots, sticks and health behaviours: a framework for documenting the complexity of financial incentive interventions to change health behaviours. *Health Psychology Review*, 8, 286-295.
- AKUNNE, A. F., BRIDGES, J. F., SANON, M. & SAUERBORN, R. 2006. Comparison of individual and group valuation of health state scenarios across communities in West Africa. *Applied Health Economics and Health Policy*, 5, 261-268.
- ALLAIS, M. 1953. Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica*, 21, 503-546.
- AMBRUS, A., GREINER, B. & PATHAK, P. 2009. Group versus individual decision-making: Is there a shift. *Institute for Advanced Study, School of Social Science Economics Working Paper*, 91.
- AMERIKS, J., CAPLIN, A., LEAHY, J. & TYLER, T. 2007. Measuring self-control problems. *American Economic Review*, 97, 966-972.
- ANDERSEN, S., COX, J. C., HARRISON, G. W., LAU, M. I., RUTSTRÖM, E. E. & SADIRAJ, V. 2011. Asset integration and attitudes to risk: theory and evidence. *Review of Economics and Statistics*.
- ANDERSEN, S., HARRISON, G. W., LAU, M. I. & RUTSTRÖM, E. E. 2006. Elicitation using multiple price list formats. *Experimental Economics*, 9, 383-405.
- ANDERSEN, S., HARRISON, G. W., LAU, M. I. & RUTSTRÖM, E. E. 2008. Eliciting risk and time preferences. *Econometrica*, 76, 583-618.
- ANDERSON, L. R. & MELLOR, J. M. 2008. Predicting health behaviors with an experimental measure of risk preference. *Journal of Health Economics*, 27, 1260-74.

References

- ARONSSON, M., HUSBERG, M., KALKAN, A., ECKARD, N. & ALWIN, J. 2015. Differences between hypothetical and experience-based value sets for EQ-5D used in Sweden: Implications for decision makers. *Scandinavian Journal of Public Health*, 43, 848-854.
- ARORA, M. & KUMARI, S. 2015. Risk taking in financial decisions as a function of age, gender: mediating role of loss aversion and regret. *International Journal of Applied Psychology*, 5, 83-89.
- 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.
- ARROW, K. J. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review*, 53, 941-973.
- ATTEMA, A.E. & LIPMAN, S.A. 2018. Decreasing impatience for health outcomes and its relation with healthy behavior. *Frontiers in Applied Mathematics and Statistics*.
- ATTEMA, A. E., BLEICHRODT, H. & L'HARIDON, O. 2018a. Measuring ambiguity preferences for health. *Health Economics*, 1-18.
- 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*, 1-9.
- ATTEMA, A. E., BLEICHRODT, H., L'HARIDON, O., PERETTI-WATEL, P. & SEROR, V. 2018b. Discounting health and money: New evidence using a more robust method. *Journal of Risk and Uncertainty*, 56, 117-140.
- ATTEMA, A. E., BLEICHRODT, H., ROHDE, K. I. M. & WAKKER, P. P. 2010. Time-tradeoff sequences for analyzing discounting and time inconsistency. *Management Science*, 56, 2015-2030.
- ATTEMA, A. E., BLEICHRODT, H. & WAKKER, P. P. 2012. A direct method for measuring discounting and QALYs more easily and reliably. *Medical Decision Making*, 32, 583-93.
- ATTEMA, A. E. & BROUWER, W. B. F. 2008. Can we fix it? Yes we can! But what? A new test of procedural invariance in TTO-measurement. *Health Economics*, 17, 877-885.
- ATTEMA, A. E. & BROUWER, W. B. F. 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. F. 2010a. On the (not so) constant proportional trade-off in TTO. *Quality of Life Research*, 19, 489-497.
- ATTEMA, A. E. & BROUWER, W. B. F. 2010b. The value of correcting values: influence and importance of correcting TTO scores for time preference. *Value in Health*, 13, 879-884.
- ATTEMA, A. E. & BROUWER, W. B. F. 2012a. Constantly proving the opposite? A test of CPTO using a broad time horizon and correcting for discounting. *Quality of Life Research*, 21, 25-34.
- ATTEMA, A. E. & BROUWER, W. B. 2012b. A test of independence of discounting from quality of life. *Journal of Health Economics*, 31, 22-34.
- 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.
- ATTEMA, A. E. & BROUWER, W. B. F. 2014. Deriving time discounting correction factors for TTO tariffs. *Health Economics*, 23, 410-425.
- 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, 1057-1065.

- ATTEMA, A. E., BROUWER, W. B.F., 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.
- AUGESTAD, L. A., RAND-HENDRIKSEN, K., KRISTIANSSEN, I. S. & STAVEM, K. 2012. Learning effects in time trade-off based valuation of EQ-5D health states. *Value in Health*, 15, 340-345.
- BAILLON, A., BLEICHRODT, H., EMIRMAHMUTOGLU, A., JASPERSEN, J. G. & PETER, R. 2020. When risk perception gets in the way: Probability weighting and underprevention. *Operations Research*.
- BALTUSSEN, R. & NIESSEN, L. 2006. Priority setting of health interventions: the need for multi-criteria decision analysis. *Cost Effectiveness and Resource Allocation*, 4, 14.
- BARON, J. & UBEL, P. A. 2001. Revising a priority list based on cost-effectiveness: The role of the prominence effect and distorted utility judgments. *Medical Decision Making*, 21, 278-287.
- BATEMAN, I., DAY, B., LOOMES, G. & SUGDEN, R. 2007a. Can ranking techniques elicit robust values? *Journal of Risk and Uncertainty*, 34, 49-66.
- BATEMAN, I., DENT, S., PETERS, E., SLOVIC, P. & STARMER, C. 2007b. The affect heuristic and the attractiveness of simple gambles. *Journal of Behavioral Decision Making*, 20, 365-380.
- BAUCELLS, M. & HEUKAMP, F. H. 2012. Probability and time trade-off. *Management Science*, 58, 831-842.
- BAUCELLS, M., WEBER, M. & WELFENS, F. 2011. Reference-Point Formation and Updating. *Management Science*, 57, 506-519.
- BENNETT, J. E., STEVENS, G. A., MATHERS, C. D., BONITA, R., REHM, J., KRUK, M. E., RILEY, L. M., DAIN, K., KENGNE, A. P. & CHALKIDOU, K. 2018. NCD Countdown 2030: worldwide trends in non-communicable disease mortality and progress towards Sustainable Development Goal target 3.4. *The Lancet*, 392, 1072-1088.
- BERNHEIM, B. D. & RANGEL, A. 2005. *Behavioral public economics: Welfare and policy analysis with non-standard decision-makers*. National Bureau of Economic Research.
- BESHEARS, J., CHOI, J. J., LAIBSON, D. & MADRIAN, B. C. 2008. How are preferences revealed? *Journal of Public Economics*, 92, 1787-1794.
- BHATIA, S. & LOOMES, G. 2017. Noisy preferences in risky choice: A cautionary note. *Psychological Review*, 124, 678.
- BHATTACHARYA, J., GARBER, A. M. & GOLDHABER-FIEBERT, J. D. 2015. *Nudges in exercise commitment contracts: a randomized trial*. National Bureau of Economic Research.
- BIRNBAUM, M. H. 2000. Decision making in the lab and on the Web. *Psychological experiments on the Internet*. Elsevier.
- BIRNBAUM, M. H. 2006. Evidence against prospect theories in gambles with positive, negative, and mixed consequences. *Journal of Economic Psychology*, 27, 737-761.
- BJÖRKMAN NYQVIST, M., CORNO, L., DE WALQUE, D. & SVENSSON, J. 2018. Incentivizing safer sexual behavior: evidence from a lottery experiment on HIV prevention. *American Economic Journal: Applied Economics*, 10, 287-314.
- BLAIS, A.-R. & WEBER, E. U. 2006. A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1.
- BLEICHRODT, H. 2001. Probability weighting in choice under risk: an empirical test. *Journal of Risk and Uncertainty*, 23, 185-198.
- BLEICHRODT, H. 2002. A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics*, 11, 447-56.

References

- 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., DOCTOR, J. & STOLK, E. A. 2005. A nonparametric elicitation of the equity-efficiency trade-off in cost-utility analysis. *Journal of Health Economics*, 24, 655-678.
- BLEICHRODT, H., DOCTOR, J. N., GAO, Y., LI, C., MEEKER, D. & WAKKER, P. P. 2019. Resolving Rabin's Paradox. *Journal of Risk and Uncertainty*, 52, 213-231.
- BLEICHRODT, H., GAO, Y. & ROHDE, K. I. M. 2016. A measurement of decreasing impatience for health and money. *Journal of Risk and Uncertainty*, 52, 213-231.
- 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. & MIYAMOTO, J. 2003. A characterization of quality-adjusted life-years under cumulative prospect theory. *Mathematics of Operations Research*, 28, 181-193.
- BLEICHRODT, H. & PINTO, J. L. 2000. A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management Science*, 46, 1485-1496.
- BLEICHRODT, H. & PINTO, J. L. 2002. Loss aversion and scale compatibility in two-attribute trade-offs. *Journal of Mathematical Psychology*, 46, 315-337.
- 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.
- BLEICHRODT, H. & PINTO PRADES, J. L. 2009. New evidence of preference reversals in health utility measurement. *Health Economics*, 18, 713-726.
- BLEICHRODT, H., SCHMIDT, U. & ZANK, H. 2009. Additive utility in prospect theory. *Management Science*, 55, 863-873.
- BLEICHRODT, H., VAN RIJN, J. & JOHANNESSON, M. 1999. Probability weighting and utility curvature in QALY-based decision making. *Journal of Mathematical Psychology*, 43, 238-260.
- BONE, J., HEY, J. & SUCKLING, J. 1999. Are groups more (or less) consistent than individuals? *Journal of Risk and Uncertainty*, 18, 63-81.
- BONTEMPO, R. N., BOTTOM, W. P. & WEBER, E. U. 1997. Cross-cultural differences in risk perception: A model-based approach. *Risk Analysis*, 17, 479-488.
- 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.
- BOWLING, A. 1995. What things are important in people's lives? A survey of the public's judgements to inform scales of health related quality of life. *Social Science & Medicine*, 41, 1447-1462.
- BRAGA, J. & STARMER, C. 2005. Preference Anomalies, Preference Elicitation and the Discovered Preference Hypothesis. *Environmental and Resource Economics*, 32, 55-89.
- BRAZIER, J., ROBERTS, J. & DEVERILL, M. 2002. The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, 21, 271-292.
- BREYER, F. & FUCHS, V. R. 1982. *Risk attitudes in health: An exploratory study*. National Bureau of Economic Research.

- BRICKMAN, P., COATES, D. & JANOFF-BULMAN, R. 1978. Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology*, 36, 917.
- BROSIG-KOCH, J., HENNIG-SCHMIDT, H., KAIRIES-SCHWARZ, N. & WIESEN, D. 2016. Using artefactual field and lab experiments to investigate how fee-for-service and capitation affect medical service provision. *Journal of Economic Behavior & Organization*, 131, 17-23.
- BROUWER, W. B. F., VAN EXEL, N. J. A. & STOLK, E. A. 2005. Acceptability of less than perfect health states. *Social Science & Medicine*, 60, 237-46.
- BROUWER, W. B. F. & VAN EXEL, N. J. A. 2005. Expectations regarding length and health related quality of life: some empirical findings. *Social Science & Medicine*, 61, 1083-1094.
- BRUHIN, A., FEHR-DUDA, H. & EPPER, T. 2010. Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, 78, 1375-1412.
- BRUNER, D. M. 2011. Multiple switching behaviour in multiple price lists. *Applied Economics Letters*, 18, 417-420.
- BRUNETTE, M., CABANTOUS, L. & COUTURE, S. 2015. Are individuals more risk and ambiguity averse in a group environment or alone? Results from an experimental study. *Theory and Decision*, 78, 357-376.
- BRYAN, G., KARLAN, D. & NELSON, S. 2010. Commitment devices. *Annual Review of Economics*, 2, 671-698.
- BUTLER, D. & LOOMES, G. 1988. Decision difficulty and imprecise preferences. *Acta Psychologica*, 68, 183-196.
- BUTLER, D. J. & LOOMES, G. C. 2007. Imprecision as an account of the preference reversal phenomenon. *American Economic Review*, 97, 277-297.
- CAMERER, C. F. & HOGARTH, R. M. 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19, 7-42.
- CARLSSON, F., MØRKBÅK, M. R. & OLSEN, S. B. 2012. The first time is the hardest: A test of ordering effects in choice experiments. *Journal of Choice Modelling*, 5, 19-37.
- CBS/RIVM 2018. *Gezondheidsenquête/leefstijlmonitor*. Den Haag.
- CHANG, S.-C., TANG, Y.-C. & LIU, Y.-J. 2016. Beyond objective knowledge: The moderating role of field dependence–independence cognition in financial decision making. *Social Behavior and Personality: an International Journal*, 44, 519-527.
- CHAPMAN, G. B. 1996. Temporal discounting and utility for health and money. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 771.
- CHAPMAN, G. B. & ELSTEIN, A. S. 1995. Valuing the future: Temporal discounting of health and money. *Medical Decision Making*, 15, 373-386.
- CHAPMAN, G. B. & SONNENBERG, F. A. 2003. *Decision making in health care: theory, psychology, and applications*, Cambridge University Press.
- COHEN, J. 1988. *Statistical power analysis for the behavioral sciences 2nd edn*. Erlbaum Associates, Hillsdale.
- COHEN, J. 2019. At over \$2 million Zolgensma is the world's most expensive therapy, yet relatively cost-effective. *Forbes*.
- COX, J. C., SADIRAJ, V., VOGT, B. & DASGUPTA, U. 2013. Is there a plausible theory for decision under risk? A dual calibration critique. *Economic Theory*, 54, 305-333.
- CRAIG, B. M., RAND, K., BAILEY, H. & STALMEIER, P. F. 2018. Quality-Adjusted Life-Years without Constant Proportionality. *Value in Health*.

References

- CURLEY, S. P., ERAKER, S. A. & YATES, J. F. 1984. An investigation of patient's reactions to therapeutic uncertainty. *Medical Decision Making*, 4, 501-511.
- DAMSCHRODER, L. J., ZIKMUND-FISHER, B. J. & UBEL, P. A. 2005. The impact of considering adaptation in health state valuation. *Social Science & Medicine*, 61, 267-277.
- DAMSCHRODER, L. J., ZIKMUND-FISHER, B. J. & UBEL, P. A. 2008. Considering adaptation in preference elicitation. *Health Psychology*, 27, 394.
- DANDURAND, F., SHULTZ, T. R. & ONISHI, K. H. 2008. Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, 40, 428-434.
- DE DREU, C. K. W., CARNEVALE, P. J. D., EMANS, B. J. M. & VAN DE VLIERT, E. 1994. Effects of Gain-Loss Frames in Negotiation: Loss Aversion, Mismatching, and Frame Adoption. *Organizational Behavior and Human Decision Processes*, 60, 90-107.
- DECI, E. L. & RYAN, R. M. 2008. Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology/Psychologie Canadienne*, 49, 182.
- DECK, C., LEE, J., REYES, J. & ROSEN, C. 2012. Risk-Taking Behavior: An Experimental Analysis of Individuals and Dyads. *Southern Economic Journal*, 79, 277-299.
- DENANT-BOEMONT, L., DIECIDUE, E. & L'HARIDON, O. 2017. Patience and time consistency in collective decisions. *Experimental Economics*, 20, 181-208.
- DEVLIN, N. J., SHAH, K. K., FENG, Y., MULHERN, B. & VAN HOUT, B. 2018. Valuing health-related quality of life: An EQ-5 D-5 L value set for England. *Health Economics*, 27, 7-22.
- DIECIDUE, E. & WAKKER, P. P. 2001. On the intuition of rank-dependent utility. *Journal of Risk and Uncertainty*, 23, 281-298.
- DOLAN, P. 1997. Modeling valuations for EuroQol health states. *Medical Care*, 35, 1095-1108.
- DOLAN, P. 2000. The measurement of health-related quality of life for use in resource allocation decisions in health care. *Handbook of Health Economics*, 1, 1723-1760.
- DOLAN, P., GUDEX, C., KIND, P. & WILLIAMS, A. 1996. The time trade-off method: results from a general population study. *Health Economics*, 5, 141-154.
- DOLAN, P., PEASGOOD, T. & WHITE, M. 2008. Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology*, 29, 94-122.
- DONKERS, B., MELENBERG, B. & VAN SOEST, A. 2001. Estimating risk attitudes using lotteries: A large sample approach. *Journal of Risk and Uncertainty*, 22, 165-195.
- DRUMMOND, M. F., SCULPHER, M. J., CLAXTON, K., STODDART, G. L. & TORRANCE, G. W. 2015. *Methods for the Economic Evaluation of Health Care Programmes*, Oxford University Press.
- EIBACH, R. P. & EHRLINGER, J. 2006. "Keep your eyes on the prize": reference points and racial differences in assessing progress toward equality. *Personality and Social Psychology Bulletin*, 32, 66-77.
- ERAKER, S. A. & SOX, H. C. 1981. Assessment of patients' preferences for therapeutic outcomes. *Medical Decision Making*, 1, 29-39.
- ESSER, J. K. 1998. Alive and well after 25 years: A review of groupthink research. *Organizational Behavior and Human Decision Processes*, 73, 116-141.
- FEENY, D. & ENG, K. 2005. A test of prospect theory. *International Journal of Technology Assessment in Health Care*, 21, 511-516.
- FEHR-DUDA, H., DE GENNARO, M. & SCHUBERT, R. 2006. Gender, Financial Risk, and Probability Weights. *Theory and Decision*, 60, 283-313.

- FISCHER, G. W., CARMON, Z., ARIELY, D. & ZAUBERMAN, G. 1999. Goal-based construction of preferences: Task goals and the prominence effect. *Management Science*, 45, 1057-1075.
- FRANCIS, L. J., BROWN, L. B. & PHILIPCHALK, R. 1992. The development of an abbreviated form of the Revised Eysenck Personality Questionnaire (EPQR-A): Its use among students in England, Canada, the USA and Australia. *Personality and Individual Differences*, 13, 443-449.
- FRASER-MACKENZIE, P., SUNG, M. C. & JOHNSON, J. E. 2014. Toward an understanding of the influence of cultural background and domain experience on the effects of risk-pricing formats on risk perception. *Risk Analysis*, 34, 1846-1869.
- GÄCHTER, S., JOHNSON, E. J. & HERRMANN, A. 2007. Individual-level loss aversion in riskless and risky choices.
- GALIZZI, M. M. 2014. What is really behavioral in behavioral health policy? And does it work? *Applied Economic Perspectives and Policy*, 36, 25-60.
- GALIZZI, M. M., MACHADO, S. R. & MINIACI, R. 2016a. Temporal stability, cross-validity, and external validity of risk preferences measures: experimental evidence from a UK representative sample.
- GALIZZI, M. M., MIRALDO, M. & STAVROPOULOU, C. 2016b. In Sickness but Not in Wealth: Field Evidence on Patients' Risk Preferences in Financial and Health Domains. *Medical Decision Making*, 36, 503-517.
- GALIZZI, M. M., MIRALDO, M., STAVROPOULOU, C. & VAN DER POL, M. M. 2016c. Doctor-patient differences in risk and time preferences: A field experiment. *Journal of health economics*, 50, 171-182.
- GALIZZI, M. M. & NAVARRO-MARTÍNEZ, D. 2018. On the external validity of social preference games: a systematic lab-field study. *Management Science*.
- GALIZZI, M. M., TAMMI, T., GODAGER, G., LINNOSMAA, I. & WIESEN, D. 2015. *Provider altruism in health economics*. National Institute for Health and Welfare.
- GALIZZI, M. M. & WIESEN, D. 2018. Behavioral experiments in health economics. *Oxford Research Encyclopedia of Economics and Finance*. Oxford: Oxford University Press.
- GERMINE, L., NAKAYAMA, K., DUCHAINE, B. C., CHABRIS, C. F., CHATTERJEE, G. & WILMER, J. B. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19, 847-857.
- GILES, E. L., ROBALINO, S., MCCOLL, E., SNIEHOTTA, F. F. & ADAMS, J. 2014. The effectiveness of financial incentives for health behaviour change: systematic review and meta-analysis. *PLoS ONE*, 9, e90347.
- GINÉ, X., KARLAN, D. & ZINMAN, J. 2010. Put your money where your butt is: a commitment contract for smoking cessation. *American Economic Journal: Applied Economics*, 2, 213-35.
- GODAGER, G. & WIESEN, D. 2013. Profit or patients' health benefit? Exploring the heterogeneity in physician altruism. *Journal of Health Economics*, 32, 1105-1116.
- GONZALEZ, R. & WU, G. 1999. On the shape of the probability weighting function. *Cognitive Psychology*, 38, 129-166.
- GOTTLIEB, D. 2012. Prospect theory, life insurance, and annuities. *The Wharton School Research Paper*.
- GREEN, C. & GERARD, K. 2009. Exploring the social value of health-care interventions: a stated preference discrete choice experiment. *Health economics*, 18, 951-976.
- GREETHER, D. M. & PLOTT, C. R. 1979. Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, 69, 623-638.

References

- HAISLEY, E., VOLPP, K. G., PELLATHY, T. & LOEWENSTEIN, G. 2012. The Impact of Alternative Incentive Schemes on Completion of Health Risk Assessments. *American Journal of Health Promotion*, 26, 184-188.
- HALEK, M. & EISENHAEUER, J. G. 2001. Demography of risk aversion. *Journal of Risk and Insurance*, 1-24.
- HALPERN, S. D., FRENCH, B., SMALL, D. S., SAULSGIVER, K., HARHAY, M. O., AUDRAIN-MCGOVERN, J., LOEWENSTEIN, G., BRENNAN, T. A., ASCH, D. A. & VOLPP, K. G. 2015. Randomized trial of four financial-incentive programs for smoking cessation. *New England Journal of Medicine*, 372, 2108-2117.
- HALPERN, S. D., KOHN, R., DORNBRAND-LO, A., METKUS, T., ASCH, D. A. & VOLPP, K. G. 2011. Lottery-based versus fixed incentives to increase clinicians' response to surveys. *Health Services Research*, 46, 1663-1674.
- HAMM, R. M. 1979. *The conditions of occurrence of the preference reversal phenomenon*. Harvard University.
- HANSEN, K. S. & ØSTERDAL, L. P. 2006. Models of quality-adjusted life years when health varies over time: survey and analysis. *Journal of Economic Surveys*, 20, 229-255.
- HARRISON, G. W., JOHNSON, E., MCINNES, M. M. & RUTSTRÖM, E. E. 2005. Temporal stability of estimates of risk aversion. *Applied Financial Economics Letters*, 1, 31-35.
- HARRISON, G. W., LAU, M. I., ROSS, D. & SWARTHOUT, J. T. 2017. Small stakes risk aversion in the laboratory: A reconsideration. *Economics Letters*, 160, 24-28.
- HARRISON, G. W. & LIST, J. A. 2004. Field experiments. *Journal of Economic Literature*, 42, 1009-1055.
- HARSANYI, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63, 309-321.
- HARTOG, J., FERRER-I-CARBONELL, A. & JONKER, N. 2002. Linking measured risk aversion to individual characteristics. *Kyklos*, 55, 3-26.
- HEINTZ, E., KROL, M. & LEVIN, L.-Å. 2013. The impact of patients' subjective life expectancy on time tradeoff valuations. *Medical Decision Making*, 33, 261-270.
- HENNIG-SCHMIDT, H., SELTEN, R. & WIESEN, D. 2011. How payment systems affect physicians' provision behaviour—an experimental investigation. *Journal of Health Economics*, 30, 637-646.
- HENRICH, J., HEINE, S. J. & NORENZAYAN, A. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61-83.
- 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.
- HERTWIG, R. & ORTMANN, A. 2001. Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 24, 383-403.
- HEY, J. D. & LOTITO, G. 2009. Naive, resolute or sophisticated? A study of dynamic decision making. *Journal of Risk and Uncertainty*, 38, 1-25.
- HIMMLER, S. F. W., PERRY-DUXBURY, M. S., VAN EXEL, N. J. A. & BROUWER, W. B. F. 2020. Willingness to pay for an early warning system for infectious diseases. *European Journal of Health Economics*, 21, 763-773.
- HOLT, C. A. & LAURY, S. K. 2002. Risk aversion and incentive effects. *American Economic Review*, 92, 1644-1655.

- HUBER, J., ARIELY, D. & FISCHER, G. 2002. Expressing preferences in a principal-agent task: A comparison of choice, rating, and matching. *Organizational Behavior and Human Decision Processes*, 87, 66-90.
- IRVINE, A., VAN DER POL, M. M. & PHIMISTER, E. 2019. A comparison of professional and private time preferences of General Practitioners. *Social Science & Medicine*, 222, 256-264.
- JACQUEMET, N. & L'HARIDON, O. 2018. *Experimental Economics*, Cambridge University Press.
- JANIS, I. L. 1972. Victims of groupthink: a psychological study of foreign-policy decisions and fiascoes.
- JONKER, M. F., ATTEMA, A. E., DONKERS, B., STOLK, E. A. & VERSTEEGH, M. M. 2017. Are health state valuations from the general public biased? A test of health state reference dependency using self-assessed health and an efficient discrete choice experiment. *Health Economics*, 26, 1534-1547.
- KAHNEMAN, D. & TVERSKY, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 263-291.
- KAIRIES-SCHWARZ, N., KOKOT, J., VOMHOF, M. & WEBLING, J. 2017. Health insurance choice and risk preferences under cumulative prospect theory—an experiment. *Journal of Economic Behavior & Organization*, 137, 374-397.
- KARIMI, M., BRAZIER, J. & PAISLEY, S. 2019. The effect of reflection and deliberation on health state values. *Value in Health*, 22, 1311-1317.
- KECK, S., DIECIDUE, E. & BUDESCU, D. V. 2014. Group decisions under ambiguity: Convergence to neutrality. *Journal of Economic Behavior & Organization*, 103, 60-71.
- KELLER, L. R., SARIN, R. K. & SOUNDERPANDIAN, J. 2007. An examination of ambiguity aversion: Are two heads better than one? *Judgment and Decision Making* (2), 5, 390-397
- KEMEL, E. & PARASCHIV, C. 2018. Deciding about human lives: an experimental measure of risk attitudes under prospect theory. *Social Choice and Welfare*, 51, 163-192.
- KIM, S.-H., AHN, J., OCK, M., SHIN, S., PARK, J., LUO, N. & JO, M.-W. 2016. The EQ-5D-5L valuation study in Korea. *Quality of Life Research*, 25, 1845-1852.
- KIMMEL, S. E., TROXEL, A. B., LOEWENSTEIN, G., BRENSINGER, C. M., JASKOWIAK, J., DOSHI, J. A., LASKIN, M. & VOLPP, K. 2012. Randomized trial of lottery-based incentives to improve warfarin adherence. *American Heart Journal*, 164, 268-274.
- KÖBBERLING, V. & WAKKER, P. P. 2005. An index of loss aversion. *Journal of Economic Theory*, 122, 119-131.
- KOOP, G. J. & JOHNSON, J. G. 2012. The use of multiple reference points in risky decision making. *Journal of Behavioral Decision Making*, 25, 49-62.
- KÖSZEGI, B. & RABIN, M. 2006. A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121, 1133-1165.
- KRABBE, P. F., ESSINK-BOT, M.-L. & BONSEL, G. J. 1996. On the equivalence of collectively and individually collected responses: standard-gamble and time-tradeoff judgments of health states. *Medical Decision Making*, 16, 120-132.
- KULLGREN, J. T., TROXEL, A. B., LOEWENSTEIN, G., NORTON, L. A., GATTO, D., TAO, Y., ZHU, J., SCHOFIELD, H., SHEA, J. A. & ASCH, D. A. 2016. A randomized controlled trial of employer matching of employees' monetary contributions to deposit contracts to promote weight loss. *American Journal of Health Promotion*, 30, 441-452.
- LAIBSON, D. 1997. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112, 443-478.
- LAZEAR, E. P. 2000. Economic imperialism. *The Quarterly Journal of Economics*, 115, 99-146.

References

- LEIDL, R. & REITMEIR, P. 2011. A value set for the EQ-5D based on experienced health states. *Pharmacoeconomics*, 29, 521-534.
- LICHTENSTEIN, S. & SLOVIC, P. 1971. Reversals of preference between bids and choices in gambling decisions. *Journal of experimental psychology*, 89, 46.
- LIPMAN, S. A. & ATTEMA, A. E. 2019. Rabin's paradox for health outcomes. *Health economics*, 28, 1064-1071.
- LIPMAN, S. A., BROUWER, W. B.F. & ATTEMA, A. E. 2019. A QALY loss is a QALY loss is a QALY loss: a note on independence of loss aversion from health states. *European Journal of Health Economics*, 20(3), 419-426
- LIPMAN, S. A., BROUWER, W. B.F. & ATTEMA, A. E. 2020. Living up to expectations: Experimental tests of subjective life expectancy as reference point in time trade-off and standard gamble. *Journal of Health Economics*, 102318.
- LIPMAN, S. A., BROUWER, W. B. F. & ATTEMA, A. E. 2019a. The Corrective Approach: Policy Implications of Recent Developments in QALY Measurement Based on Prospect Theory. *Value in Health*, 22, 816-821.
- LIPMAN, S. A., BROUWER, W. B. F. & ATTEMA, A. E. 2019b. QALYs without bias? Non-parametric correction of time trade-off and standard gamble weights based on prospect theory. *Health Economics*, 28, 843-854.
- 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.
- LOEWENSTEIN, G., O'DONOGHUE, T. & RABIN, M. 2003. Projection bias in predicting future utility. *Quarterly Journal of Economics*, 118, 1209-1248.
- LOEWENSTEIN, G. & PRELEC, D. 1991. Negative time preference. *American Economic Review*, 81, 347-352.
- LOOMES, G. & SUGDEN, R. 1982. Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, 92, 805-824.
- MACKEIGAN, L. D., GAFNI, A. & O'BRIEN, B. J. 2003. Double discounting of QALYs. *Health Economics*, 12, 165-169.
- MANKIW, N. G. 2020. *Principles of economics*, Cengage Learning.
- MANTZARI, E., VOGT, F., SHELMT, I., WEI, Y., HIGGINS, J. P. & MARTEAU, T. M. 2015. Personal financial incentives for changing habitual health-related behaviors: A systematic review and meta-analysis. *Preventive Medicine*, 75, 75-85.
- 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.
- MASLOW, A. H. 1943. A theory of human motivation. *Psychological Review*, 50, 370-396.
- MASSIN, S., NEBOUT, A. & VENTELOU, B. 2018. Predicting medical practices using various risk attitude measures. *European Journal of Health Economics*, 19, 843-860.
- MCCABE, C., BRAZIER, J., GILKS, P., TSUCHIYA, A., ROBERTS, J., O'HAGAN, A. & STEVENS, K. 2006. Using rank data to estimate health state utility models. *Journal of Health Economics*, 25, 418-431.
- MCINTOSH, C. N., GORBER, S. C., BERNIER, J. & BERTHELOT, J.-M. 2007. Eliciting Canadian population preferences for health states using the Classification and Measurement System of Functional Health (CLAMES). *Chronic Disease Canada*, 28, 29-41.

- MENZEL, P., DOLAN, P., RICHARDSON, J. & OLSEN, J. A. 2002. The role of adaptation to disability and disease in health state valuation: a preliminary normative analysis. *Social Science & Medicine*, 55, 2149-2158.
- MITCHELL, M. S., GOODMAN, J. M., ALTER, D. A., JOHN, L. K., OH, P. I., PAKOSH, M. T. & FAULKNER, G. E. 2013. Financial incentives for exercise adherence in adults: systematic review and meta-analysis. *American Journal of Preventive Medicine*, 45, 658-667.
- MIYAMOTO, J. M. & ERAKER, S. A. 1988. A multiplicative model of the utility of survival duration and health quality. *Journal of Experimental Psychology: General*, 117, 3.
- 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.
- MIYAMOTO, J. M., WAKKER, P. P., BLEICHRODT, H. & PETERS, H. J. 1998. The zero-condition: a simplifying assumption in QALY measurement and multiattribute utility. *Management Science*, 44, 839-849.
- NICE 2018. *Guide to the processes of technology appraisal*. London.
- NORMAN, R., CRONIN, P. & VINEY, R. 2013. A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. *Applied Health Economics and Health Policy*, 11, 287-298.
- NORMAN, R., MULHERN, B. & VINEY, R. 2016. The impact of different DCE-based approaches when anchoring utility scores. *Pharmacoeconomics*, 34, 805-814.
- 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.
- NOUSSAIR, C. & WU, P. 2006. Risk tolerance in the present and the future: An experimental study. *Managerial and Decision Economics*, 27, 401-412.
- OLIVER, A. 2003a. The internal consistency of the standard gamble: tests after adjusting for prospect theory. *Journal of Health Economics*, 22, 659-674.
- OLIVER, A. 2003b. A quantitative and qualitative test of the Allais paradox using health outcomes. *Journal of Economic Psychology*, 24, 35-48.
- OLIVER, A. 2006. Further evidence of preference reversals: choice, valuation and ranking over distributions of life expectancy. *Journal of Health Economics*, 25, 803-820.
- OLIVER, A. 2013a. *Behavioural public policy*, Cambridge University Press.
- OLIVER, A. 2013b. Testing the rate of preference reversal in personal and social decision-making. *Journal of Health Economics*, 32, 1250-1257.
- OLIVER, A. 2015. Nudging, shoving, and budging: Behavioural economic-informed policy. *Public Administration*, 93, 700-714.
- OLIVER, A. 2018. Your money and your life: Risk attitudes over gains and losses. *Journal of Risk and Uncertainty*, 57, 29-50.
- OLIVER, A. & SUNSTEIN, C. 2019. Does size matter? The Allais paradox and preference reversals with varying outcome magnitudes. *Journal of Behavioral and Experimental Economics*, 78, 45-60.
- OPPE, M., DEVLIN, N. J., VAN HOUT, B., KRABBE, P. F. & DE CHARRO, F. 2014. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health*, 17, 445-453.
- PACHUR, T., HERTWIG, R. & WOLKEWITZ, R. 2014. The affect gap in risky choice: Affect-rich outcomes attenuate attention to probability information. *Decision*, 1, 64.

References

- PACHUR, T., SUTER, R. S. & HERTWIG, R. 2017. How the twain can meet: Prospect theory and models of heuristics in risky choice. *Cognitive Psychology*, 93, 44-73.
- PALOYO, A. R., REICHERT, A. R., REUSS-BORST, M. & TAUCHMANN, H. 2015. Who responds to financial incentives for weight loss? Evidence from a randomized controlled trial. *Social Science & Medicine*, 145, 44-52.
- PATEL, M. S., ASCH, D. A., ROSIN, R., SMALL, D. S., BELLAMY, S. L., HEUER, J., SPROAT, S., HYSON, C., HAFF, N. & LEE, S. M. 2016. Framing financial incentives to increase physical activity among overweight and obese adults: a randomized, controlled trial. *Annals of Internal Medicine*, 164, 385-394.
- PAYNE, J. W. 2005. It is whether you win or lose: The importance of the overall probabilities of winning or losing in risky choice. *Journal of Risk and Uncertainty*, 30, 5-19.
- PEER, E., EGELMAN, S., HARBACH, M., MALKIN, N., MATHUR, A. & FRIK, A. 2019. Nudge Me Right: Personalizing Online Nudges to People's Decision-Making Styles.
- PÉNTEK, M., BRODSZKY, V., GULÁCSI, Á. L., HAJDÚ, O., EXEL, N.J.A., BROUWER, W.B.F. & GULÁCSI, L. 2014. Subjective expectations regarding length and health-related quality of life in Hungary: results from an empirical investigation. *Health Expectations*, 17, 696-709.
- PERPIÑÁN, J. M. A., MARTÍNEZ, F. I. S., PÉREZ, J. E. M. & MARTÍNEZ, I. M. 2009. *Debiasing eq-5d tariffs. New estimations of the Spanish EQ-5D value set under nonexpected utility.* Centro de Estudios Andaluces.
- PICKARD, A. S., LAW, E. H., JIANG, R., PULLENAYEGUM, E., SHAW, J. W., XIE, F., OPPE, M., BOYE, K. S., CHAPMAN, R. H. & GONG, C. L. 2019. United States Valuation of EQ-5D-5L Health States Using an International Protocol. *Value in Health*, 22, 931-941.
- PINTO-PRADES, J.-L. & ABELLAN-PERPIÑÁN, J.-M. 2012. When normative and descriptive diverge: how to bridge the difference. *Social Choice and Welfare*, 38, 569-584.
- PINTO-PRADES, J. L., SÁNCHEZ-MARTÍNEZ, F. I., ABELLÁN-PERPIÑÁN, J. M. & MARTÍNEZ-PÉREZ, J. E. 2018. Reducing preference reversals: The role of preference imprecision and nontransparent methods. *Health Economics*, 27, 1230-1246.
- PLISKIN, J. S., SHEPARD, D. S. & WEINSTEIN, M. C. 1980. Utility functions for life years and health status. *Operations Research*, 28, 206-224.
- RABIN, M. 2000. Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, 68, 1281-1292.
- RABIN, M. & THALER, R. H. 2001. Anomalies: risk aversion. *Journal of Economic Perspectives*, 15, 219-232.
- RANGANATHAN, M. & LAGARDE, M. 2012. Promoting healthy behaviours and improving health outcomes in low and middle income countries: a review of the impact of conditional cash transfer programmes. *Preventive Medicine*, 55, S95-S105.
- RAPPANGE, D. R., BROUWER, W. B. F. & VAN EXEL, N. J. A. 2016. A long life in good health: subjective expectations regarding length and future health-related quality of life. *European Journal of Health Economics*, 17, 577-589.
- 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.
- REDELMEIER, D. A. & HELLER, D. N. 1993. Time preference in medical decision making and cost-effectiveness analysis. *Medical Decision Making*, 13, 212-217.
- REILLY, R. J. 1982. Preference reversal: Further evidence and some suggested modifications in experimental design. *American Economic Review*, 72, 576-584.

- RIVA, G., TERUZZI, T. & ANOLLI, L. 2003. The use of the internet in psychological research: comparison of online and offline questionnaires. *CyberPsychology & Behavior*, 6, 73-80.
- ROBINSON, A., DOLAN, P. & WILLIAMS, A. 1997. Valuing health status using VAS and TTO: what lies behind the numbers? *Social Science & Medicine*, 45, 1289-1297.
- ROBINSON, A., LOOMES, G. & JONES-LEE, M. 2001. Visual analog scales, standard gambles, and relative risk aversion. *Medical Decision Making*, 21, 17-27.
- ROBINSON, A. & SPENCER, A. 2006. Exploring challenges to TTO utilities: valuing states worse than dead. *Health Economics*, 15, 393-402.
- ROBINSON, S. & BRYAN, S. 2013. Does the process of deliberation change individuals' health state valuations? An exploratory study using the person trade-off technique. *Value in Health*, 16, 806-813.
- ROCKENBACH, B., SADRIEH, A. & MATHAUSCHEK, B. 2007. Teams take the better risks. *Journal of Economic Behavior & Organization*, 63, 412-422.
- ROHDE, K. I. M. 2010. The hyperbolic factor: A measure of time inconsistency. *Journal of Risk and Uncertainty*, 41, 125-140.
- ROHDE, K. I. M. 2019. Measuring Decreasing and Increasing Impatience. *Management Science*, 65, 1700-1716.
- ROSEN, A. B., TSAI, J. S. & DOWNS, S. M. 2003. Variations in risk attitude across race, gender, and education. *Medical Decision Making*, 23, 511-517.
- RUBINSTEIN, A. 2012. *Lecture notes in microeconomic theory: the economic agent*, Princeton University Press.
- RYAN, M., BATE, A., EASTMOND, C. & LUDBROOK, A. 2001. Use of discrete choice experiments to elicit preferences. *BMJ Quality & Safety*, 10, i55-i60.
- SACKETT, D. L. & TORRANCE, G. W. 1978. The utility of different health states as perceived by the general public. *Journal of chronic diseases*, 31, 697-704.
- SAVAGE, L. J. 1954. *The Foundations of Statistics*, New York, Wiley.
- SCHRAM, A. 2005. Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, 12, 225-237.
- SEIDL, C. 2002. Preference reversal. *Journal of Economic Surveys*, 16, 621-655.
- SESTON, E. M., ASHCROFT, D. M. & GRIFFITHS, C. E. 2007. Balancing the benefits and risks of drug treatment: a stated-preference, discrete choice experiment with patients with psoriasis. *Archives of Dermatology*, 143, 1175-1179.
- SHAFIR, E. 2013. *The behavioral foundations of public policy*, Princeton University Press.
- SHIBA, S. & SHIMIZU, K. 2019. Does time inconsistency differ between gain and loss? An intra-personal comparison using a non-parametric elicitation method. *Theory and Decision*.
- SHUPP, R. S. & WILLIAMS, A. W. 2007. Risk preference differentials of small groups and individuals. *The Economic Journal*, 118, 258-283.
- SLOVIC, P. 1995. The construction of preference. *American Psychologist*, 50, 364.
- SPRANCA, M., MINSK, E. & BARON, J. 1991. Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, 27, 76-105.
- STALMEIER, P. F. & BEZEMBINDER, T. G. 1999. The discrepancy between risky and riskless utilities: a matter of framing? *Medical Decision Making*, 19, 435-447.

References

- STALMEIER, P. F., WAKKER, P. P. & BEZEMBINDER, T. G. 1997. Preference reversals: Violations of unidimensional procedure invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 1196.
- 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.
- STEIN, K., RATCLIFFE, J., ROUND, A., MILNE, R. & BRAZIER, J. E. 2006. Impact of discussion on preferences elicited in a group setting. *Health and Quality of Life Outcomes*, 4, 22.
- 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.
- STOLK, E. A., LUDWIG, K., RAND, K., VAN HOUT, B. & RAMOS-GOÑI, J. M. 2019. Overview, update, and lessons learned from the international EQ-5D-5L valuation work: Version 2 of the EQ-5D-5L Valuation Protocol. *Value in Health*, 22, 23-30.
- STOLK, E. A., OPPE, M., SCALONE, L. & KRABBE, P. F. 2010. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. *Value in Health*, 13, 1005-1013.
- STROHACKER, K., GALARRAGA, O. & WILLIAMS, D. M. 2013. The impact of incentives on exercise behavior: a systematic review of randomized controlled trials. *Annals of Behavioral Medicine*, 48, 92-99.
- SUTER, R. S., PACHUR, T. & HERTWIG, R. 2016. How affect shapes risky choice: Distorted probability weighting versus probability neglect. *Journal of Behavioral Decision Making*, 29, 437-449.
- SUTER, R. S., PACHUR, T., HERTWIG, R., ENDESTAD, T. & BIELE, G. 2015. The neural basis of risky choice with affective outcomes. *PLoS ONE*, 10, e0122475.
- SUTHERLAND, H. J., LLEWELLYN-THOMAS, H., BOYD, N. F. & TILL, J. E. 1982. Attitudes Toward Quality of Survival: The Concept of "Maximal Endurable Time". *Medical Decision Making*, 2, 299-309.
- TANGNEY, J. P., BOONE, A. L. & BAUMEISTER, R. F. 2018. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Self-Regulation and Self-Control*. Routledge.
- THALER, R. H. 2000. From homo economicus to homo sapiens. *Journal of Economic Perspectives*, 14, 133-141.
- THALER, R. H. & SUNSTEIN, C. R. 2009. *Nudge: Improving decisions about health, wealth, and happiness*, Penguin.
- TOPLAK, M. E., WEST, R. F. & STANOVICH, K. E. 2011. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition*, 39, 1275.
- TORRANCE, G. W. 1976. Toward a utility theory foundation for health status index models. *Health Services Research*, 11, 349.
- TORRANCE, G. W. 1987. Utility approach to measuring health-related quality of life. *Journal of Chronic Diseases*, 40, 593-600.
- 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. *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.

- TVERSKY, A., SLOVIC, P. & KAHNEMAN, D. 1990. The causes of preference reversal. *The American Economic Review*, 204-217.
- TVERSKY, A. & THALER, R. H. 1990. Anomalies: preference reversals. *Journal of Economic Perspectives*, 4, 201-211.
- VAN DE WETERING, E., STOLK, E. A., VAN EXEL, N. J. A. & BROUWER, W. B. F. 2013. Balancing equity and efficiency in the Dutch basic benefits package using the principle of proportional shortfall. *European Journal of Health Economics*, 14, 107-115.
- VAN DER POL, M. M. & ROUX, L. 2005. Time preference bias in time trade-off. *European Journal of Health Economics*, 6, 107-111.
- VAN DER POL, M. M. & RUGGERI, M. 2008. Is risk attitude outcome specific within the health domain? *Journal of Health Economics*, 27, 706-717.
- VAN DER POL, M. M. & CAIRNS, J. A. 2000. Negative and zero time preference for health. *Health Economics*, 9, 171-175.
- VAN DER SWALUW, K., LAMBOOIJ, M. S., MATHIJSSSEN, J. J., SCHIPPER, M., ZEELLENBERG, M., BERKHOUT, S., POLDER, J. J. & PRAST, H. M. 2018. Commitment Lotteries Promote Physical Activity Among Overweight Adults—A Cluster Randomized Trial. *Annals of Behavioral Medicine*, 52, 342-351.
- VAN NOOTEN, F. & BROUWER, W. B. F. 2004. The influence of subjective expectations about length and quality of life on time trade-off answers. *Health Economics*, 13, 819-823.
- VAN NOOTEN, F., KOOLMAN, X. & BROUWER, W. B. F. 2009. The influence of subjective life expectancy on health state valuations using a 10 year TTO. *Health Economics*, 18, 549-558.
- VAN NOOTEN, F., KOOLMAN, X., BUSSCHBACH, J. & BROUWER, W. B. F. 2014. Thirty down, only ten to go?! Awareness and influence of a 10-year time frame in TTO. *Quality of Life Research*, 23, 377-384.
- VAN OSCH, S. M. & STIGGELBOUT, A. M. 2008. The construction of standard gamble utilities. *Health Economics*, 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. *Medical 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. *Medical Decision Making*, 24, 511-7.
- VAN ROSSUM, C., BUURMA, E., VENNEMANN, F., BEUKERS, M., DRIJVERS, J. & OCKÉ, M. 2017. Voedselconsumptie in 2012-2014 vergeleken met de Richtlijnen goede voeding 2015.
- VAN UFFELEN, J. G., WONG, J., CHAU, J. Y., VAN DER PLOEG, H. P., RIPHAGEN, I., GILSON, N. D., BURTON, N. W., HEALY, G. N., THORP, A. A. & CLARK, B. K. 2010. Occupational sitting and health risks: a systematic review. *American journal of preventive medicine*, 39, 379-388.
- VAN WILGENBURG, K. 2018. Beliefs, Preferences and Health Insurance Behavior.
- VAN WINNSEN, K., VAN KLEEF, R. & VAN DE VEN, W. 2016. Potential determinants of deductible uptake in health insurance: How to increase uptake in The Netherlands? *European Journal of Health Economics*, 17, 1059-1072.
- VASHISTHA, A., CUTRELL, E. & THIES, W. Increasing the reach of snowball sampling: The impact of fixed versus lottery incentives. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2015. ACM, 1359-1363.

References

- 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. & BROUWER, W. B. F. 2016. Patient and general public preferences for health states: a call to reconsider current guidelines. *Social Science & Medicine*, 165, 66-74.
- 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.
- VOLPP, K. G., JOHN, L. K., TROXEL, A. B., NORTON, L., FASSBENDER, J. & LOEWENSTEIN, G. 2008. Financial incentive-based approaches for weight loss: a randomized trial. *JAMA*, 300, 2631-2637.
- VON GAUDECKER, H.-M., VAN SOEST, A. & WENGSTRÖM, E. 2008. Selection and mode effects in risk preference elicitation experiments.
- VON NEUMANN, J. & MORGENSTERN, O. 1944. *Theory of games and economic behavior* New York, Wiley.
- WAKKER, P. P. & DENEFFE, D. 1996. Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science*, 42, 1131-1150.
- WAKKER, P. P. & STIGGELBOUT, A. 1995. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Medical Decision Making*, 15, 180-186.
- WAKKER, P. P. 2008. Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, 17, 1329-1344.
- WAKKER, P. P. 2010. *Prospect theory: For risk and ambiguity*, Cambridge university press.
- WALTERS, S. J. & BRAZIER, J. E. 2005. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Quality of Life research*, 14, 1523-1532.
- WANG, X. T. & JOHNSON, J. G. 2012. A tri-reference point theory of decision making under risk. *Journal of Experimental Psychology: General*, 141, 743.
- WEBER, E. U., BLAIS, A. R. & BETZ, N. E. 2002. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15, 263-290.
- WEBER, E. U. & HSEE, C. 1998. Cross-cultural differences in risk perception, but cross-cultural similarities in attitudes towards perceived risk. *Management Science*, 44, 1205-1217.
- WHO 1948. Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948. http://www.who.int/governance/eb/who_constitution_en.pdf.
- WHO 2009. Global health risks : mortality and burden of disease attributable to selected major risks. Geneva: World Health Organization.
- WHO 2018. Public spending on health: a closer look at global trends. World Health Organization.
- WOERNER, A. 2018. Overcoming Time Inconsistency with a Matched Bet: Theory and Evidence from Exercising.
- WONG, E. L., RAMOS-GOÑI, J. M., CHEUNG, A. W., WONG, A. Y. & RIVERO-ARIAS, O. 2018. Assessing the use of a feedback module to model EQ-5D-5L health states values in Hong Kong. *The Patient-Patient-Centered Outcomes Research*, 11, 235-247.
- WOUTERS, S. 2016. *Absolutely Relative: on the value of health outcomes*.

- WOUTERS, S., VAN EXEL, N. J. A. , ROHDE, K. I. M. & BROUWER, W. B. F. 2015. Are all health gains equally important? An exploration of acceptable health as a reference point in health care priority setting. *Health & Quality of Life Outcomes*, 13, 79.
- XIE, F., PULLENAYEGUM, E., GAEBEL, K., BANSBACK, N., BRYAN, S., OHINMAA, A., POISSANT, L. & JOHNSON, J. A. 2016. A time trade-off-derived value set of the EQ-5D-5L for Canada. *Medical Care*, 54, 98.
- XIE, F., PULLENAYEGUM, E., GAEBEL, K., OPPE, M. & KRABBE, P. F. 2014. Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best-worst scaling? *European Journal of Health Economics*, 15, 281-288.
- YECHIAM, E. & HOCHMAN, G. 2013. Losses as modulators of attention: review and analysis of the unique effects of losses over gains. *Psychological Bulletin*, 139, 497.
- ZHANG, J. & CASARI, M. 2012. How groups reach agreement in risky choices: an experiment. *Economic Inquiry*, 50, 502-515.
- ZIN 2015. Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg. *Diemen: Zorginstituut Nederland*.

About the author

Stefan Lipman (1992) was Honours student in Social Psychology (B.Sc.) at Utrecht University, and completed a research master in Social and Health Psychology (M.Sc.) also in Utrecht. During his studies, he developed a profound passion for teaching and an interest in interdisciplinary research. Having worked with social and health psychologists, legal scholars and veterinarians, he continued his interdisciplinary research effort by pursuing a Ph.D. at Erasmus School of Health Policy & Management.

His research focuses (in the broadest sense) on applying insights from behavioral and health economics to understand decisions about health. As such, he studied a wide array of topics in his research, such as: intertemporal choice for health, exercise behavior, risk attitudes for health, loss aversion, collective decision-making for health, preference reversals, reference-points for health, and even motivations to engage in pest control. Special focuses of his work are measurement of risk and time preferences for health outcomes (e.g. using prospect theory) and applying such measures in health state valuation.

After finishing his dissertation, Stefan was a visiting scholar at the Max Planck Institute for Human Development, and continued working at Erasmus School of Health Policy & Management (ESHPM) as assistant professor involved in several collaborative research and teaching projects.

