"For every complex problem, there is a solution that is simple, neat, and wrong."

- Henry Louis Mencken
Influence of construct-irrelevant factors and effects of methodological choices on EQ-5D health state valuation

Liv Ariane Augestad and Kim Rand-Hendriksen

Thesis submitted for the degree of Philosophiae Doctor (PhD)
at the Institute of Clinical Medicine
University of Oslo
February 2012

Research conducted at
Health Services Research Centre, Akershus University Hospital
and
Department of Health Management and Health Economics, University of Oslo
"For every complex problem, there is a solution that is simple, neat, and wrong."
- Henry Louis Mencken
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table of contents</td>
<td>1</td>
</tr>
<tr>
<td>Preface</td>
<td>7</td>
</tr>
<tr>
<td>Authorship</td>
<td>11</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>13</td>
</tr>
<tr>
<td>List of frequently used abbreviations</td>
<td>19</td>
</tr>
<tr>
<td>Other important points about terminology</td>
<td>23</td>
</tr>
<tr>
<td>List of papers</td>
<td>25</td>
</tr>
<tr>
<td>1 Background</td>
<td>27</td>
</tr>
<tr>
<td>1.1 Health economic analyses and health-related quality of life</td>
<td>27</td>
</tr>
<tr>
<td>1.2 Quality-adjusted life years</td>
<td>28</td>
</tr>
<tr>
<td>1.2.1 Indirect HRQoL measurement/MAUIs</td>
<td>28</td>
</tr>
<tr>
<td>1.2.2 ‘Utility’, ‘preferences’, or ‘values’</td>
<td>29</td>
</tr>
<tr>
<td>1.3 EQ-5D</td>
<td>30</td>
</tr>
<tr>
<td>1.3.1 Brief history and the instrument</td>
<td>30</td>
</tr>
<tr>
<td>1.3.2 Health states and value sets/tariffs</td>
<td>31</td>
</tr>
<tr>
<td>1.3.2.1 Group of reference</td>
<td>34</td>
</tr>
<tr>
<td>1.3.2.2 Conventional methods and existing national tariffs</td>
<td>34</td>
</tr>
<tr>
<td>1.4 Time trade-off</td>
<td>36</td>
</tr>
<tr>
<td>1.4.1 General concepts and brief history</td>
<td>36</td>
</tr>
<tr>
<td>1.4.2 MVH-style TTO</td>
<td>38</td>
</tr>
<tr>
<td>1.4.2.1 Transformation of states considered as WTD</td>
<td>42</td>
</tr>
<tr>
<td>1.4.3 Lead-time TTO</td>
<td>44</td>
</tr>
<tr>
<td>1.5 Other valuation methods</td>
<td>45</td>
</tr>
</tbody>
</table>
1.5.1 Ranking and discrete choice experiments 45
1.5.2 Visual analogue scale 45
1.5.3 Comparisons of valuation methods 47

1.6 Construct relevance and procedural invariance 47

1.6.1 Heuristics and cognitive biases 48
  1.6.1.1 Framing effects 49
  1.6.1.2 Anchoring 49
  1.6.1.3 Focusing 50

1.6.2 Communicative implicatures and the cooperative principle 50

1.6.3 Invoking death 52

2 Objectives 55

3 Methods 59

3.1 External datasets 60
  3.1.1 UK EQ-5D valuation study 61
  3.1.2 US EQ-5D valuation study 62
  3.1.3 Medical expenditure survey panel 63

3.2 Valuation experiments 64
  3.2.1 Sampling of respondents 64
  3.2.2 Structure of the surveys 65
    3.2.2.1 Demographics 65
    3.2.2.2 Valuation 65
    3.2.2.3 Attitudes toward euthanasia 66
  3.2.3 TTO variants 68
    3.2.3.1 MVH TTO 68
3.2.3.2 LT-TTO
3.2.3.3 TTO5
3.2.4 Respondent groups
3.2.5 Exclusion criteria
3.3 Analyses by paper
  3.3.1 Paper I: Collapsed TTO
  3.3.2 Paper II: WTD transformation
  3.3.3 Paper III: Learning effects in TTO
  3.3.4 Paper IV: Ranking vs. TTO
  3.3.5 Paper V: Experienced and hypothetical health
  3.3.6 Paper VI: Attitudes toward euthanasia
  3.3.7 Paper VII: Anchoring bias in TTO
3.4 Permissions and ethics
4 main results
  4.1.1 Paper I: Collapsed TTO
  4.1.2 Paper II: WTD transformation
  4.1.3 Paper III: Learning effects in TTO
  4.1.4 Paper IV: Ranking vs. TTO
  4.1.5 Paper V: Experienced and hypothetical health
  4.1.6 Paper VI: Attitudes toward euthanasia
  4.1.7 Paper VII: Anchoring bias in TTO
5 discussion
  5.1 Discussion of methods
    5.1.1 External datasets (papers I-V)
    5.1.2 Valuation experiments (papers VI and VII)
5.1.2.1 Mode of administration
5.1.2.2 Exclusion criteria
5.1.3 Regression modeling
5.1.4 Discussion of methods by paper
  5.1.4.1 Paper I: Collapsed TTO
  5.1.4.2 Paper II: WTD transformation
  5.1.4.3 Paper III: Learning effects in TTO
  5.1.4.4 Paper IV: Ranking vs. TTO
  5.1.4.5 Paper V: Experienced and hypothetical health
  5.1.4.6 Paper VI: Attitudes toward euthanasia
  5.1.4.7 Paper VII: Anchoring bias in TTO
5.2 General discussion
  5.2.1 Influence of research choices and construct-irrelevant factors
    5.2.1.1 WTD transformation – paper II
    5.2.1.2 Anchoring – paper VII
    5.2.1.3 Learning effects – paper III
    5.2.1.4 Attitudes toward euthanasia – paper VI
    5.2.1.5 Collapsed TTO – paper I
  5.2.2 Absolute values
  5.2.3 The initial choice task
  5.2.4 Worse than death values
  5.2.5 The value of ‘death’
  5.2.6 Rank inconsistencies
  5.2.7 Tariffs as measures of the preferences of the researchers
  5.2.8 Experience or prediction
<table>
<thead>
<tr>
<th></th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Implications and future perspectives</td>
<td>129</td>
</tr>
<tr>
<td>7</td>
<td>Conclusions</td>
<td>131</td>
</tr>
<tr>
<td>8</td>
<td>Errata</td>
<td>133</td>
</tr>
<tr>
<td>9</td>
<td>References</td>
<td>135</td>
</tr>
<tr>
<td>10</td>
<td>Appendix: Survey questions</td>
<td>151</td>
</tr>
<tr>
<td>11</td>
<td>Papers I-VII</td>
<td>159</td>
</tr>
</tbody>
</table>
PREFACE

Things did not go according to plan. This may be true for most PhD projects, but probably not to this extent. To help making sense of this thesis, we will quickly recap the story of what we were supposed to do, why we did not do so, and what we ended up doing.

The QALY-project (QALY being short for quality-adjusted life year), as it was informally referred to at the Research Centre, was off to a magnificent start: The two project leaders, Knut Stavem and Ivar Sønbø Kristiansen, had applied for funding from both the Norwegian Research Council and the Eastern Norway Regional Health Authorities. Both applications were granted. Thus, when the two of us were hired, the project had well-defined goals (the creation of Norwegian EQ-5D and 15D value sets/tariffs), and the financial resources required for to perform large national surveys.

The closest thing to a gold standard in EQ-5D valuation was to have a large, nationally representative respondent sample value hypothetical health states described with the EQ-5D descriptive system using the time trade-off (TTO) method. Informally, TTO is a method for determining what proportion of the remaining life time a respondent would be willing to relinquish in order to get from an impaired state of health to perfect health. To us, the concept of a trade-off between length of life and quality of life had face validity. That is, until we read the interview guide used in the US EQ-5D valuation study. The method starts by having the respondent choose between a life of 10 years in the target state (health state to be valued) and a life of zero years in perfect health, i.e. immediate death. While this might be sound from the perspective of economic theory, the psychological implications of having respondents repeatedly consider presented health states in relation to immediate death seemed daunting.
The subsequent method was different for states considered by the respondent to be better than death (BTD) and states considered to be worse than death (WTD). We found the WTD procedure far too complicated to merit the assumption that respondents would be able to respond sensibly. Therefore, less than a month into the project period, we were starting to question whether we could, in good faith, perform a TTO-based EQ-5D valuation study.

Our reservations regarding the TTO method led to serious discussions with our supervisors. Understandably, they were interested in reaching the stated goals of the project, i.e. creating Norwegian value sets for the EQ-5D and the 15D. We argued that it would be premature and against our integrity to perform a costly valuation study based on methods we considered lacking in validity. Unable to reach any quick consensus on the issue of EQ-5D, we agreed to plan the survey necessary for creating a value algorithm for the 15D while writing up our initial findings on the EQ-5D and looking for ways to make an EQ-5D valuation study acceptable.

After approximately one year of writing papers based on data from previous EQ-5D valuation studies, studying literature, debating, and arguing, we agreed not to perform an EQ-5D valuation study. Instead, we would collect data aiming to shed light on some of the worrying areas relating to validity of TTO valuation.

The altered pathway of the project had consequences for the methods and topics included in this thesis. We started out trying to learn the methods used in valuation, and found ourselves looking for ways to use existing data and statistical methods typical to valuation studies to substantiate our concerns. Several of our objections to TTO are based on assumptions of a psychological nature: that discomfort induced by repeated invocation of “immediate death” has unwanted implications for how people respond, that the starting point and routing of the TTO directly influences resulting values, that the WTD task is too complicated, etc. However, readers expecting to find qualitative analyses or analyses of the
methods based primarily on psychological theory will be disappointed. A strength and a weakness of this thesis is use of methods familiar to most people in the field in new ways to illustrate issues related to how valuation studies have been performed. Hopefully, by improving understanding of these issues, this thesis may be a contribution to the improvement of future health state valuation methods.
AUTHORSHIP

From the beginning of this project, it seems that the conventional ways of doing things have not suited us. In the same spirit, even though convention would dictate separate theses, we opted to make it a co-production. Since we started out, the path of the project and the papers has been determined by constant throwing ideas back-and-forth, discussing of hypotheses, thinking aloud, mutual questioning, and some arguing. We therefore considered that any division of the results of our joint work would be artificial; there would have been substantial overlaps, but most importantly, it would have required some extent of arbitrary division of content.

We have shared the same office most of the project period, and are used to sitting together in a huge office, about 40 cm away from each other, typing. Usually, we have been typing on different papers (for which we have been first authors), however, producing this thesis has been the most intertwined part of our project. Splitting up this thesis into who has written what is thus anything but a simple task. Most sections (including this one) have been written and rewritten several times, in some instances by one of the authors and in the next instance by the other.

Due to the requirements of the PhD, in that individual contributions should be identifiable, we have tried to determine sections for which one of us has greater responsibility. The primary division of labor has been related to the papers: while both of us have contributed at all steps in the writing process, the first author performed the analyses, wrote most of the paper, and had the final word in cases of disagreement. Thus, LAA takes most of the credit for papers II, III, and VI, while KRH takes most credit for papers I, IV, and V. We conceived of the idea behind paper VII very early in the project. After the data collection was complete, we were approached by Carl Haakon Samuelsen, who was looking for good subjects for his Master’s thesis. He performed analyses and wrote his thesis based on our
anchoring data, the central parts of which we have jointly adapted to become paper VII.

With the exception of the papers, we have added a solid gray line to the left of the parts of the texts primarily attributable to LAA (look to the left)

and a dotted gray line to the left of parts primarily attributable to KRH.

However, several sections were such that we could not divide them between us, for instance the description of objectives, implications, and conclusions. The same applies to the longest single section of the discussion, regarding the value of ‘death’. However, for simplicity, we have determined that LAA should answer to the section on death, while KRH answers to paper VII on anchoring.
ACKNOWLEDGEMENTS

Roughly four years ago, we met in the hallways of MIR – the temporary contraption on stilts, named for its resemblance to the not-so-picturesque space station, that housed the Research Centre at AHUS prior to the finish of the new hospital wing. Years of fun and frustration – and suddenly, we were finished. Fortunately, we were not without support.

First of all, we would like to thank our supervisors Knut Stavem and Ivar Sønbø Kristiansen, who made the whole project possible. They came up with the idea, applied for the necessary funding and were kind (or smart?) enough to employ us. The project took its shape through friendly guidance, discussion and ...resistance by our supervisors. Yes, we’ve had our disagreements, particularly in the first year, but we’re quite happy with the end result, and hope you are, too. With their extensive experience in the publishing game, Knut and Ivar certainly went over and beyond the call of duty with regard to reading numerous versions of our papers of varying quality. We hope the papers have improved somewhat over time. Thank you for fruitful collaboration 😊

The project would not have been possible without funding. It so happens that Knut and Ivar sent two applications: one to the Norwegian Research Council, and one to the South-Eastern Norway Regional Health Authority. Both were granted, which is why there were two of us. Dear Research Council and Regional Health Authority: we have much to thank you for. Food on the table, clothes on our backs, trips to Greece, France, Canada and the US, and future prospects involving experimental testing of the long-term consequences of coffee abuse. Thank you for your support!

The academic committee that considered this thesis should be applauded. Not only did they reach the right conclusion, but they were willing to delve into a thesis that is roughly twice the normal size. Thus, they’ve performed twice the
work, knowing that they’d only be invited to one dinner. This self-sacrifice is appreciated by us.

We would like to thank Harri Sintonen for helping us understand the internal workings of the 15D instrument. The thesis proper ended up being focused on EQ-5D, but a substantial amount of work underway has been related to 15D and comparing 15D with EQ-5D.

When the thesis was finally done, work was just about to begin for the faculty administration: Natalia Andronova and Mari Nenningsland Edvardsen have been a lot more helpful than we had any reason to anticipate or expect. If everyone working in administrations had been equipped with such an attitude, the world would probably look quite different.

We would like to express our gratitude to HØKH, for creating a friendly work environment, for instance by letting us have the biggest and brightest corner office (which after our departure lodges a Professor) and putting up with our crazy ideas, like organizing a “whining-seminar” for all PhD students, and wasting a day of work for the whole research centre on an April’s fool hoax. We definitely have to mention the wonderful administrative staff (Nina Viksløkken Ødegård, Karin Vassbakk and Reidun Skårerhøgda) who helped us keep track of the budget (very complicated stuff), and kept us from starving by ensuring that we got our salary, as well as helping with innumerable strange and unusual requests. And coffee.

Our bosses at HØKH, Hilde Lurås (the Big Boss) and Fredrik Dahl (the Not Quite As Big, But Probably Quite Strong Boss (we believe this is how he would like us to think of him)) have been stern, funny, helpful, nagging, organizing and kind.

Our colleagues have been great, and have contributed to us wanting to go all the way out to AHUS, and taking less vacation time than we really should have. Yes, we blame that on you guys. They have enabled our penchant for empty calories
by joining us for walks to the local food store, helped us keep our laughing muscles exercised, and become more than just colleagues – our friends 😊 We have been fortunate to have people around who realize that, while we’ve had a lot of fun by ourselves in the corner office, we’ve had a lot more fun with visitors. A mixture of academic debate and general nonsense has helped us get through the more tedious parts of the scientific process.

Some colleagues we’ve badgered more than others, and they deserve specific mention. Mathias: we’re skipping you for a different section. Deal with it. Lene Berge Holm is probably the one who’s workday has been most disrupted by our activities. Whenever we’ve heard the sound of high heels in the distance, we’ve turned away from our screens and expected relief in the form of laughter and smiles. Fredrik Gregersen (AKA Fredrikito or Little-Fredrik (in spite of the fact that he is much taller than both of us, possibly even taller than Big Fredrik (Strong Fredrik. Sorry.))) started at HØKH about half-way through our PhD, and has also worked at HELED, thereby becoming a sort of double colleague. Thanks for banter on boats, statistics, advisors, datasets, etc. Pål Gulbrandsen has become some sort of half-mentor in the arts of political incorrectness. He likes to meddle with everything and everyone, which is a good thing. We’ve enjoyed your company and input.

At HELED (an outdated acronym for the department of health management and health economics at the University of Oslo), we have been enjoying the company of a group going under the name the Journal Club, where we’ve been granted insight into the strange and wondrous world of economists in their natural habitat. Since our project has tangented on behavioral economy, this has been a very interesting and useful experience. Outside of the journal club proper, we have likely overstayed our welcome in the office of our eminent colleague Arna Desser, who has been more than patient with our lack of understanding of economic theory, and has had the grace to debate our points all the same. We
sincerely hope to work with you in the future. We’ve also really enjoyed the company of Emily Burger, who has been constantly supportive and constructive.

Liv:

I would like to thank my family, especially my parents, for supporting an academic interest while I grew up. And to my mother - when I actually went for an academic degree, she did not forget about the basics, like eating vegetables, staying warm on my feet and the occasional worry about my ability to survive colds. All this while enduring a daughter periodically too absent-minded to call her and when we finally spoke, all kinds of ranting about the difficulty of writing up papers, stupid reviewers etc. Je parie que tu es contente que j’ai fini ma thèse!

Mathias, my colleague, best friend, and partner in romance. People wonder why we spend so much time together, and I guess the explanation is in the former sentence. There are a lot of things I would like to say to you, but for this time (we don’t want to embarrass people) let’s focus on how you helped the making of this thesis. You know most details of the (dis)QALY project and that is why your belief in the project, even before we had any published papers, has been especially appreciated. You have been supportive at all hours of the day and participated in rather lengthy discussions on rather gloomy topics (often involving death) over dinner, in the car, at work, at home and between sessions of cross words.

Dear Kim, it has been a more than a pleasure to work with you over the past few years. Playing idea-ping-pong with you must certainly be one of the best and funniest kinds of brain-exercise, making work with you more like a hobby. Discussions and conversations have not been limited to the project, but involved everything from politics, psychology, nuclear power, and why it’s a good idea to be on facebook. From time to time we’ve had our differences, but for the most part, and for the important parts we’ve been in agreement - which is a good thing.
when one decides to write a joint thesis. It makes me happy in other ways as well: I consider you to have a pretty great mind, and since great minds think alike...;) Thank you for participating in this and future projects with me, both as a friend and colleague, it is greatly appreciated.

Kim:
I would like to thank my parents, who have been endlessly supportive and genuinely interested all the way from beginning to end. To their relief, my father managed to finish his PhD-thesis more than a year ahead of me. Their recent experiences with the seemingly endless process of getting everything ready have been a real help.

Mentors and advisors on my Cand.Psychol-thesis, Hilde Eileen Nafstad and Rolv Mikkel Blakar have been instrumental in me considering going for a PhD, and supportive, even when I ended up doing a different project than the one we started on years ago.

My colleague and friend Hanne Brorson has, as always, been supportive of even my strangest projects, and has pushed me to excel. A big thanks to several of my friends – for bearing with me and listening patiently when I’ve ranted about this thesis, which has been sometimes tedious, definitely long-lasting, and very much on topics for those of specific interests: Gry Aavik (the section on communicative implicatures came about partly from our discussions), Kristina Koller (thanks for the support, particularly when I was just about to get started), Tor Erlend Pedersen (bear in mind that I’ve listened to your antics as well ;-)), Ann Christin Cappelen (thanks for getting my mind on other things), Tuva Øktedalen (We started out dancing, then became psychologists, and now researchers. What should we do next?), and, of course, my brothers Ole and Morten.
Mathias Barra deserves special mention: Your running off with Liv clearly contributed negatively to the project for a while. She became more distant, and both less present at, and less concentrated on, work. However, her mood, which has typically been quite good, improved further, and soon, we were back on (off?) track. Debates with you turned out to be both interesting and challenging. Not to mention that you turned out to be a great guy whom I now consider a good friend. Thank you.

Finally, I would like to thank Liv. Of all the people I could have ended up with as my partner in this project, I was exceedingly lucky to end up with you. It’s been interesting, it’s been fun, and I’ve looked forward to just about every day of work with you. If I didn’t have someone to confide in when things looked most bleak, I would have quit. For close to four years, now, we’ve been working side by side, and it’s been great. So great that those few times when we’ve been at different places, work has been boring. I am glad you agreed to make this a joint thesis, because it really reflects how we’ve worked, and I see your hand in all of it, even the parts written by me. I actually think that this specific text may be the only text related to this project that I have written with absolutely no input from you. Which feels strange. I greatly respect you as a colleague, and look forward to continue working with you in the future. But most of all, I appreciate your friendship, and long after we’ve stopped writing papers side by side, I dearly hope to retain it.

Thank you.

Oslo, August 3rd 2012

Liv and Kim
# List of Frequently Used Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATE</td>
<td>Attitudes toward euthanasia.</td>
</tr>
<tr>
<td>BTD</td>
<td>Better than death. Health state considered to be better than death by a valuation method, resulting in a positive value. Also described in parts of the literature as SBD – state better than death.</td>
</tr>
<tr>
<td>cTTO</td>
<td>Collapsed TTO. A procedure that collapses TTO values on to four values: Unwilling to trade (1), better than death (.5), equal to death (0), and worse than death (-.5).</td>
</tr>
<tr>
<td>DCE</td>
<td>Discrete choice experiments. Valuation method that asks respondents to choose the best of two alternatives.</td>
</tr>
<tr>
<td>EH</td>
<td>Experienced health. Relating to how respondents who are in an impaired health state value their own health.</td>
</tr>
<tr>
<td>ETD</td>
<td>Equal to death. Health state considered to be equally bad as immediate death, giving it a value of zero.</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>Five-dimensional generic questionnaire/instrument used as an indirect measure of HRQoL. EQ refers to EuroQol, the consortium behind the development of the instrument. The questionnaire also includes the EQ-VAS, a thermometer-like VAS used to measure HRQoL.</td>
</tr>
<tr>
<td>fTTO</td>
<td>Full TTO. In contrast to the cTTO (see above), the TTO values are not collapsed.</td>
</tr>
<tr>
<td>GLS</td>
<td>Generalized least-squares regression.</td>
</tr>
</tbody>
</table>
HH  Hypothetical health. In contrast to EH (see above), relating to how the general population value health states they have not experienced.

IRT  Item response theory.

LT-TTO  Lead-time TTO. A variant of TTO where time in perfect health is added before the life in the target state.

MCID  Minimal Clinically Important Difference. Estimate of the minimum magnitude of change on an instrument score that represents a noticeable improvement or worsening of health in a clinical setting.

MEPS  Medical expenditure survey panel. A set of large-scale surveys performed in the US to measure the cost, usage, and availability of health services.

MVH  Measuring and Valuing Health. The name of the project within which the UK EQ-5D valuation study was performed.

MVH TTO  Variant of TTO used in the US and UK valuation studies.

OLS  Ordinary least-squares regression.

QALY  Quality adjusted life year. A measure of gain used in economic evaluation of health interventions.

SBD  State better than death. See SWD and WTD.

SG  Standard Gamble. Health state valuation method that is a direct implementation of Von Neumann-Morgenstern’s expected utility theory. Precursor to TTO.

SWD  State worse than death. In the papers on learning effects (paper III) and transformation methods (paper II), SWD and SBD were used
instead of WTD and BTD because previous papers in the journals in question had used those abbreviations. See WTD and BTD.

**TTO** Time trade-off. Health state valuation method.

**TTOS** Variant of MVH TTO where the starting point was shifted from immediate death to 5 years in perfect health.

**UK** United Kingdom.

**US** United States.

**UTT** Unwilling to trade. In TTO, this describes a situation in which the respondent is not willing to trade away any life-time to go from being in the target state to being in perfect health. The health state is then given a value of 1, equal to perfect health.

**VAS** Visual analogue scale. A psychometric response scale. People are asked to rate the perceived magnitude of the construct of interest along a vertical or horizontal line with descriptors beyond the end points. Also referred to in the literature as rating scale (RS) or category rating (CR).

**WTD** Worse than death. Used to describe a health state that has been considered worse than death by a valuation method, giving it a value below zero. Also described in parts of the literature as SWD – state worse than death.

**WTP** Willingness to pay. Monetary valuation method for goods, including health.
OTHER IMPORTANT POINTS ABOUT TERMINOLOGY

Utility/values:

Much of the literature on health state valuation refers to Von Neumann-Morgenstern’s expected utility theory. A relatively direct implementation of this theory is the standard gamble (SG) valuation method. Whether or not values derived using other valuation methods can be regarded as utilities is an ongoing debate that is not considered to be essential or relevant to this thesis. We have therefore avoided using the word ‘utility’ in this thesis and have referred to the results of valuations as ‘values’, ‘TTO values’, or ‘VAS values’.

VAS, EQ-VAS, EQ-5D VAS, rating scale, category rating (scale):

In the literature on EQ-5D, all these terms are used to refer to the VAS included in the EQ-5D questionnaire. Unless stated otherwise, we will use VAS to refer to this in this thesis.

Patrick/Dolan/non-linear/non-monotonic/UK/MVH transformation:

The method used to transform WTD TTO values in the UK valuation study and most subsequent TTO-based valuation studies. It is referred to by all these names in the literature. We refer to it as the ‘Patrick transformation’ in the thesis and as the ‘UK transformation’ in paper II.

Torrance/linear/monotonic/US transformation:

The method used to transform WTD TTO values in the US valuation study. It is referred to by all these names in the literature. We refer to it as the ‘Torrance transformation’ in this thesis and as the ‘US transformation’ in paper II.
Numerals and numbers:

As a rule of thumb, numbers below 10 are written out as words. However, in this thesis we frequently present numbers directly related to scale values on TTO, e.g., as in the number of years in perfect health in the TTO task. In order to draw attention to the numbers’ relationship with scale values, and since such scale values need not be integers, we have chosen to present these using digits.


**List of Papers**


**Paper IV:** Rand-Hendriksen K, Augestad LA. Time trade-off and ranking exercises are sensitive to different dimensions of EQ-5D health states. Submitted.


1 BACKGROUND

1.1 HEALTH ECONOMIC ANALYSES AND HEALTH-RELATED QUALITY OF LIFE

Publicly funded health care systems aim, *inter alia*, to improve the health of target populations\(^1\). Given that such health care systems have limited budgets, it is imperative that economic evaluations of health care programs be undertaken to determine how funds should be allocated to maximize health improvement\(^2\). Publicly funded health care systems may be concerned with other aims, *e.g.* equity\(^3\). However, this thesis pertains to how health benefits are quantified for subsequent use in economic evaluations of health care programs.

Reduced mortality is a crucial benefit of health care, and for life-extending interventions, the number of individuals saved per unit of cost may be enough information to guide decision makers. However, health care increasingly has other primary goals that are less tangible and cover a range of different dimensions. For instance, the most important benefit of hip arthroplasty for osteoarthritis patients is reduced pain and increased mobility, while the health benefit of antidepressant drugs is reduced depression. To enable economic evaluation of health care programs across different types of patient groups and diagnoses, improvements related to different dimensions of health must be quantified using the same metric\(^4,5\). We commonly assume that interventions improve the health-related quality of life (HRQoL) of patients; if pain, mobility and depression did not have an impact on quality of life, the use of hip arthroplasty and antidepressants would not be justified (if we were to disregard the potential life prolonging effect of these interventions). In other words, the impact on quality of life of health care interventions can function as a common metric for assessing the health benefits of different health interventions.
1.2 QUALITY-ADJUSTED LIFE YEARS

The quality-adjusted life year (QALY) is a measure especially developed for and used in health economics\(^6\). It integrates the two main benefits that health care aims to increase: reduced mortality and increased HRQoL. To calculate QALYs, a period of time is multiplied by a weight or value that corresponds to the utility or the HRQoL associated with a specific health state, i.e. quantity of health is the integer of this weight over time\(^6\). The weight is anchored to 1, indicating perfect health, and 0, indicating a health state that is equivalent to death. Thus, one QALY equals for instance one year in perfect health or two years in a health state weighted to .5. The difference in average QALYs between, for example, a group of non-treated and a group of treated patients represents the QALY gain, or the health benefit associated with a treatment. In economic evaluations, QALY gain is typically divided by cost to render cost per QALY. Cost per QALY estimates can be compared directly for different interventions, or the incremental ratio of cost to QALY-gain over existing treatment options can be compared to an absolute incremental cost-effectiveness ratio (ICER) threshold\(^7,8\). Since the QALY measure combines both quality and quantity of life in a single number, it allows for the comparison of interventions that are life-prolonging, that improve quality of life, or both\(^9\). Assessing mortality is a matter of counting; one is usually not in doubt about whether a patient is dead. The challenge is how to measure HRQoL\(^9,10\): consider the contrast between evaluating whether some-one is dead (yes or no) and considering whether a terrible health state is equivalent to death. Measurement of HRQoL poses numerous normative, ethical, and methodological problems, some of which we shall address in this thesis.

1.2.1 INDIRECT HRQoL MEASUREMENT/MAUIs

Different methods are available to measure HRQoL (valuation methods), but directly eliciting HRQoL values from individual patients is considered time-consuming and complicated. Instead, indirect methods are often applied, using
multi-attribute utility instruments (MAUIs). Patients report on their current health by filling out MAUI questionnaires describing their health along a set of several dimensions. Each possible composite health state in the questionnaire is associated with a pre-specified value. These values are usually obtained through valuation studies in which the general population values a selection of target health states, and then an algorithm is developed that estimates the HRQoL value of all possible descriptions of health in the MAUI.

1.2.2 ‘Utility’, ‘Preferences’, or ‘Values’

‘Utility’, in everyday language, is a rather nonspecific term for “the capacity for being useful for some purpose”\textsuperscript{11}. In economics, ‘utility’ usually refers to a measure of the total satisfaction that a consumer receives from consuming a good or service. In this sense, QALYs cannot be a measure of utility \textit{per se}, but under certain assumptions about the value used to weight each year, QALYs represent a number of life years weighted by an index of utility\textsuperscript{12}. Thus, the debate on terminology refers to the quality weight, not the QALY.

In health economics, ‘utility’, both as a concept and as a unit of measurement, is often equated with ‘utility’ as described by Von-Neumann-Morgenstern’s expected utility theory (EUT)\textsuperscript{12}. This theory is a normative description of rational decision making when an outcome is uncertain\textsuperscript{13}. Since the standard gamble (SG) is a valuation method that directly implements EUT, SG is considered by many to be the preferred scaling method and has been characterized as a ‘gold standard’ in health state valuation\textsuperscript{14,15}. Several authors have criticized the definition of health state utilities as ‘SG utilities’, because individuals violate the axioms of EUT, and because disagreements exists about the appropriateness of including risk attitudes in societal decision making processes\textsuperscript{3,12}.

In the health economics literature, the terms ‘utilities’, ‘preferences’, ‘HRQoL’, and ‘values’ are used interchangeably when referring to the weights used in QALY
calculation. There is disagreement about the definitions of these terms, and the definition of the constructs they represent may have consequences for how they should be measured. For instance, some researchers argue that ‘preferences’ should be reserved for values elicited by choice-based procedures\(^{16}\), while others do not feel that a choice or a trade is necessary to express a ‘preference’\(^ {17}\).

How the target constructs are described is important. However, while the field of health economics would benefit from greater consensus regarding terminology, that particular debate is beyond the scope of this thesis. To avoid confusion and misunderstanding, we will predominantly use the terms ‘value’ and ‘health state value’, sometimes using the specific elicitation method as a prefix for clarification. With the term ‘value’, we adhere to Froberg and Kane’s definition of ‘utility’ or ‘preference’: “levels of subjective satisfaction, distress or desirability that people associate with a particular health state.”\(^ {18}\)

In our published papers, we have sometimes used ‘utilities’ or ‘preferences’; at the time of writing, we were not completely aware of the controversies surrounding this terminology, and the reviewers often had opinions regarding the terms used.

### 1.3 EQ-5D

#### 1.3.1 Brief History and the Instrument

The first meeting of the interdisciplinary EuroQol group, in 1987, aimed to develop a MAUI that would cover the most important dimensions of HRQoL covering a wide range of severity, yet be short and easy enough for self-completion, for instance in postal surveys\(^ {19}\). The resulting MAUI, the EQ-5D, has become the instrument most frequently used to collect HRQoL data for economic evaluations\(^ {20,21}\).

The revised guidelines issued by the UK National Institute of Clinical Excellence (NICE) for methods of technology appraisal stated an explicit preference for the
EQ-5D as a source of HRQoL weights for the calculation of QALYs and noted that the use of any other measure must be empirically justified\textsuperscript{22,23}. The EQ-5D description system has five dimensions of health: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has three levels of severity: no problem (level 1), some problem (level 2) and severe/extreme problem (level 3). Figures 1 and 2 display the EQ-5D self-rating questionnaire and the associated VAS.

1.3.2 **HEALTH STATES AND VALUE SETS/TARIFFS**

When a respondent fills out the EQ-5D questionnaire, the result is a health profile, or EQ-5D health state (‘health state’ for short). Such health states are conventionally described using a five-digit index, where each digit represents the level of functioning within the respective dimension in the previously presented order. Thus, health states 11111 and 33333 represent the best and the worst EQ-5D health states, respectively.
Figure 1 - The EQ-5D self-rating questionnaire

By placing a tick in one box in each group below, please indicate which statements best describe your own health state today.

**Mobility**
- I have no problems in walking about
- I have some problems in walking about
- I am confined to bed

**Self-Care**
- I have no problems with self-care
- I have some problems washing or dressing myself
- I am unable to wash or dress myself

**Usual Activities (e.g. work, study, housework, family or leisure activities)**
- I have no problems with performing my usual activities
- I have some problems with performing my usual activities
- I am unable to perform my usual activities

**Pain/Discomfort**
- I have no pain or discomfort
- I have moderate pain or discomfort
- I have extreme pain or discomfort

**Anxiety/Depression**
- I am not anxious or depressed
- I am moderately anxious or depressed
- I am extremely anxious or depressed
Figure 2 - The EQ-5D self-rating visual analogue scale

To help people say how good or bad a health state is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked 100 and the worst state you can imagine is marked 0.

We would like you to indicate on this scale how good or bad your own health is today, in your opinion. Please do this by drawing a line from the box below to whichever point on the scale indicates how good or bad your health state is today.


1.3.2.1 GROUP OF REFERENCE

When the objective is to determine how good or bad different health states are, one may pose questions about the subject to people with reduced health, their caretakers, medical experts, or the general public. Within the current paradigm, values for health states described by instruments like the EQ-5D are usually intended to reflect the preferences of the general population in the country in which the instrument is to be used\(^\text{15}\). Typically, respondents from the general population are prompted to imagine being in a health state described by the EQ-5D or similar instruments, and to measure how good or bad they believe such a state would be, using one of several valuation methods. The primary competing view has been that health states should be based on patients’ perceptions of their own health\(^\text{24–27}\). When comparing values from the general population and values from patients, patients have usually assigned higher values to their own health states than the corresponding values assigned by respondents from the general population\(^\text{26,28–30}\).

The debate about whether to use patient or general population values reflects a more general debate in economics regarding the use of what Kahneman described as ‘decision utility’, versus the use of ‘experienced utility’\(^\text{31,32}\). ‘Decision utility’ refers to the ‘wantability’ of the valued concepts\(^\text{33}\), while ‘experienced utility’ refers to how good or bad things are while experienced. The use of decision utility has been the norm in most branches of economics since the early 1900’s but is now being challenged, particularly by proponents of the use of general quality of life, or even happiness, as a goal of economic activity\(^\text{34–38}\).

1.3.2.2 CONVENTIONAL METHODS AND EXISTING NATIONAL TARIFFS

Although several other methods have been developed to value EQ-5D health states, most EQ-5D valuation studies have been conducted in the tradition of the seminal Measuring and Valuing Health (MVH) study performed in the UK in 1993\(^\text{39}\). In the UK study, respondents were first asked to rank a set of 12 health
states + perfect health (state 11111), “unconscious”, and “immediate death”, from subjective best to worst. Then, they were asked to rate the same health states, using a visual analogue scale (VAS) similar to the one used in self-ratings in the EQ-5D questionnaire. Third, respondents were asked to rate the same 12 states in random order, using time trade-off (TTO), with a protocol we will refer to as MVH-style TTO. In total, 42 of the 243 possible EQ-5D health states were valued by different respondents, and regression modeling was used to generate a value algorithm that applied to all possibilities. This procedure has become the de facto gold standard for EQ-5D valuation\textsuperscript{40,41}. Table 1 lists all published national EQ-5D valuation studies by year of data collection. Four of the countries listed have tariffs based only on Visual Analogue Scale (VAS). All 15 other countries have TTO-based tariffs as their recommended or official tariffs. Valuation methods, including VAS, TTO in general, and the MVH-style TTO protocol will be described in more detail in subsequent subsections.
Table 1 – published EQ-5D tariffs

<table>
<thead>
<tr>
<th>Country</th>
<th>Year of collection</th>
<th>Final sample (n)</th>
<th>Recommended version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finland</td>
<td>1992</td>
<td>928 VAS</td>
<td></td>
</tr>
<tr>
<td>UK</td>
<td>1993</td>
<td>2997 TTO</td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>1996-1997</td>
<td>975 TTO</td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td>1997-1998</td>
<td>339 TTO</td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td>1998</td>
<td>543 TTO</td>
<td></td>
</tr>
<tr>
<td>New Zealand</td>
<td>1999</td>
<td>396 VAS</td>
<td></td>
</tr>
<tr>
<td>Denmark</td>
<td>1999-2000</td>
<td>1332 TTO</td>
<td></td>
</tr>
<tr>
<td>Slovenia</td>
<td>2000</td>
<td>370 VAS</td>
<td></td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>2000</td>
<td>2384 TTO</td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>2001</td>
<td>548 VAS</td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>2002</td>
<td>3773 TTO</td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>2003</td>
<td>298 TTO</td>
<td></td>
</tr>
<tr>
<td>Argentina</td>
<td>2003-2004</td>
<td>611 TTO</td>
<td></td>
</tr>
<tr>
<td>South Korea</td>
<td>2007</td>
<td>1264 TTO</td>
<td></td>
</tr>
<tr>
<td>Thailand</td>
<td>2007</td>
<td>1409 TTO</td>
<td></td>
</tr>
<tr>
<td>Poland</td>
<td>2008</td>
<td>305 TTO</td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>2008</td>
<td>443 TTO</td>
<td></td>
</tr>
<tr>
<td>Chile</td>
<td>2008</td>
<td>1967 TTO</td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>2009#</td>
<td>417 TTO</td>
<td></td>
</tr>
</tbody>
</table>

* Most TTO-based valuation studies used VAS to familiarize respondents with valuation. Several of these have not published VAS-based tariffs.

# Timing of data collection confirmed through email correspondence with Richard Norman February the 1st 2012.

1.4 TIME TRADE-OFF

1.4.1 GENERAL CONCEPTS AND BRIEF HISTORY

Time trade-off is a class of valuation methods in which the common denominator is trade-offs between length and (health-related) quality of life. The aim of each valuation task is to find the point of preferential indifference between a certain amount of life-time in reduced health (t) and shorter life-time in full health (x).

The value of the reduced health state is the proportion of time in full health at equivalence, relative to the time in reduced health, i.e. x/t. The QALY is an
integrated measure of time and quality of life. Thus, the TTO method may be seen as a quite direct operationalization of the QALY.\textsuperscript{6,59}

The TTO method was suggested by Torrance\textsuperscript{60} as an alternative to standard gamble (SG), which previously had been the standard for measuring health state values. The major difference between the TTO and the SG is that the latter method measures values under conditions of uncertainty, which may be appropriate for individual decision making, where the outcome for the patient actually involves a risk, while TTO measures values under conditions of certainty, which may be more appropriate for societal values\textsuperscript{3,12,15}. A comparison of SG, TTO, and VAS suggested that TTO was the best method of measuring health state values, because it was somewhat easier for respondents to perform and had slightly higher reliability than SG, while the sample mean values for health states with TTO were similar to those obtained with SG\textsuperscript{61}.

In all variations of TTO, respondents are faced with a series of tasks, each of which is presented as choice between two hypothetical lives, usually referred to as life A and life B. For each task, the respondent is asked to determine whether he/she would prefer to live in life A and then die or live in life B and then die. Alternatively, the respondent may indicate that the two lives are perceived as equally good/bad. The lives are described with different lengths of time in the target state or perfect health and are altered based on the respondents’ choices until the respondent is indifferent between the two. Assigning a value of 1 to perfect health and 0 to death then allows calculation of the value of the target state.

Originally, the TTO method developed by Torrance\textsuperscript{60,61} did not allow eliciting values for health states considered worse than death (WTD). However, in 1978, Rosser and Kind\textsuperscript{62} questioned the assumption that ‘death’ was the worst possible outcome, since suicide and requests to withdraw from lifesaving treatments take
Empirical studies confirmed the existence of WTD health states\textsuperscript{63}, and Torrance suggested a WTD procedure to elicit values with the TTO\textsuperscript{64}.

It is important to note that TTO can be administered in numerous different ways. We will start by describing the protocol used in the UK MVH EQ-5D valuation study\textsuperscript{39} which has served as a reference protocol for other valuation studies of EQ-5D health states. The UK valuation study will be described in more detail in section 3.1.1. After presenting MVH TTO, we will describe the other variants used in this thesis.

### 1.4.2 MVH-STYLE TTO

In MVH-style TTO, respondents may describe presented health states as being better than death (BTD), equal to death (ETD), or worse than death (WTD). Within the QALY framework, death is assigned a value of 0, meaning that health states considered to be WTD are assigned negative values. In the MVH protocol, different methods are used to value states considered to be BTD and states considered to be WTD. To determine whether a presented state should be valued using the BTD or the WTD method, the initial choice task in the MVH protocol asks respondents to choose between life A described as 0 years in perfect health, i.e., immediate death, and life B described as 10 years in the target state.

If the respondent considers the two lives equally good/bad, the valuation is over for that health state, and it is assigned a TTO value of 0.
If the respondent prefers life B, the target state is considered BTD, and life A is set to 5 years in perfect health (Figure 3). If the respondent still prefers life B, life A is increased by one-year increments until the respondent either prefers life A or finds the two lives equally good/bad. If the respondent prefers life A at 5 years of perfect health, the length of life A is reduced with one-year increments. Valuation is finished when the respondent reaches a preferential equilibrium between the two lives or when the respondent has stated his/her preference with life A set to something and 6 months. If the respondent still prefers one life over the other with a resolution of 6 months, the point of equilibrium is assumed to be found midway between the two alternatives for which preference is reversed. Figure 4 displays the full routing of MVH TTO. If $t$ is used to denote years in perfect health and $x$ is used to denote years in the target state, the TTO value for the target state is set to $t / x$. In the BTD task, $x$ is always 10 years, meaning that the TTO-value $u$ is...
For states considered to be WTD in the initial choice task, the iterative comparison aims at detecting the point of equivalence between life B described as ‘immediate death’ and a composite life A consisting of one part in target state and the remainder in perfect health, such that the total life length is 10 years. In the first WTD choice task, life A is described as 5 years in the target state followed by 5 years in perfect health, followed by death (Figure 5). Conceptually, the respondent is asked to consider if he/she would accept living in the target state for \( x \) years if compensated by \( t=10-x \) years of perfect health, when the other alternative is immediate death. The length of time in target state/perfect health is altered by increments of 1 year, followed by a \( \frac{1}{2} \) year correction if necessary, until preferential equilibrium is reached.

With death assigned a value of 0 and perfect health a value of 1 in accordance with the QALY-regime, we get the equation

\[
0 = t * 1 - x * u = t - u(10 - t)
\]

Solved for \( u \)

\[
u = \frac{-t}{10 - t}
\]
Figure 4 - The routing of MVH-style TTO

Better than death.
Life A is 10 years in target state (t)
Life B varies depending on the respondent's choices:
- Prefer Life A:
- Life A and life B are equal:
- Prefer Life B:

Life A: 10 years in target health state (t)
Life B: 0 year in perfect health (p)

Worse than death
Life B constant at 0 years
Life A varies depending on the respondent's choices:
- Prefer Life A:
- Life A and life B are equal:
- Prefer Life B:
1.4.2.1 Transformation of States Considered as WTD

The BTD scores have an upper boundary of 1, while there is no theoretical lower boundary for WTD values. In line with this, the implementation of Torrance’s suggestion for eliciting WTD values requires no theoretical lower boundary. However, it follows from the method that the lowest possible value depends on the smallest amount of time respondents are allowed to trade.

In the MVH study the smallest tradable quantity was 3 months. The lowest value is reached if the respondent prefers immediate death to a life A presented as 6 months in the target state followed by 9 years and 6 months in perfect health. The point of preferential equilibrium is then assumed to be at 3 months target state, followed by 9 years, 9 months in perfect health, which gives a lowest possible TTO value of
Without transformation of WTD values, two thirds of the valued health states in the UK EQ-5D valuation study ended up with negative mean values. The authors of the UK study concluded that respondents could not have understood the WTD task as it was intended. They therefore chose to transform negative TTO-values to a range of -1 to 0 prior to aggregation. The assumption was that respondents understood the WTD task as having the same scale properties as the BTD task, i.e. that the WTD scale is an interval scale. This implies that respondents only paid attention to the years in perfect health or the years in reduced health, not both concepts simultaneously. They presented no empirical or theoretical evidence for this assumption. However, there is ample evidence that people are generally bad at judgment tasks requiring calculations, and that judgments involving fractions are among the most difficult.

If \( u' \) denotes transformed TTO values, the transformation, suggested by Patrick but referred to variously as the Dolan transformation, the UK transformation, and the non-linear transformation was

\[
    u = \frac{-t}{10 - t} = \frac{-9.75}{10 - 9.75} = \frac{-9.75}{.25} = -39
\]

The researchers behind the newer US valuation study (described in more detail in the section 3.1.2), which was modeled on experience from the preceding UK valuation study, chose to use a different transformation method suggested by Torrance. The Torrance transformation method has also been referred to as the monotonic transformation and the US transformation. If \( u' \) denotes transformed TTO values, the Torrance transformation was
Transformation of negative values has been criticized as lacking theoretical and empirical support, and the two different methods necessarily lead to different mean values. With the exception of the lowest possible value in the MVH protocol, the Torrance transformation results in values closer to 0 than the Patrick transformation. Several other approaches have been suggested for handling the problems created by the magnitude of WTD values from this TTO protocol, including truncating negative values, using of circular regression, using medians instead of means, or using a TTO protocol that allows positive and negative values to be elicited with one continuous method.

1.4.3 LEAD-TIME TTO

To overcome some of the problems related to negative values in MVH-style TTO, Devlin and colleagues suggested the use of lead-time TTO (hereafter ‘LT-TTO’) as a way to allow respondents to value WTD and BTD states with one continuous method.

LT-TTO resembles the BTD task in MVH-style TTO. However, where life B in MVH TTO is presented as 10 years in the target state, LT-TTO adds a period of perfect health before time in the target state. This initial period is referred to as lead-time. If a respondent prefers life A when it is shorter than the lead-time in life B, he/she indirectly indicates that the target state is WTD. LT-TTO is still experimental, and several aspects of it have yet to be agreed upon, including length of lead-time and time in target state, what to do if respondents “use up” all available lead-time, where along the scale the task should start, and what routing procedure should be used.
1.5 OTHER VALUATION METHODS

1.5.1 RANKING AND DISCRETE CHOICE EXPERIMENTS

A relatively simple (at least to understand) valuation method has respondents rank presented health states from subjective best to subjective worst. In EQ-5D valuation studies based on the MVH protocol, ranking has been used as a precursor to other valuation tasks.

Discrete choice experiments are even simpler than ranking: respondents are presented with two (or sometimes three) alternatives, in this case EQ-5D health states, and are asked to determine which one they prefer. Repeated a great number of times with different sets of alternatives, these experiments generate preference data that enable calculation of how good/bad the alternatives are in relation to one another through the use of Rasch models or related models from the field of item response theory (IRT). Rank data has not typically been used to generate tariffs, but this is possible: ranking can be conceptualized as ordering sets of discrete choices, meaning that the methods used to create tariffs from DCE data can be used on rank data. Work is currently underway in the EuroQol group to combine TTO and DCE for valuation purposes.

1.5.2 VISUAL ANALOGUE SCALE

A visual analogue scale (VAS) is a psychometric response scale intended to measure attitudes, feelings, characteristics, and other mental magnitudes that are assumed to range across a continuum and that cannot be measured directly. For instance, pain can be conceptualized as ranging from no pain to extreme pain, without any apparent discrete jumps.

In its simplest form, a VAS consists of a line (horizontal or vertical) with descriptors beyond each endpoint. The respondent is instructed to indicate on the line where the construct of interest intersects, representing its perceived...
magnitude on the assumed underlying continuum. Numerous variations exist: vertical/horizontal, of different lengths, with or without written/numeric descriptors placed along the line, with one or more crossing lines, etc. The appearance of the VAS has some impact on how people respond. For instance, horizontal scales have been reported to lead to distributions that are closer to normal than those obtained with vertical scales.

Relevant to this thesis is the variant usually used in relation to the EQ-5D, often referred to as the EQ-5D VAS, EQ-VAS, or thermometer scale, that is part of the EQ-5D self-rating form (Figure 2). It is defined as a 20 cm long vertical VAS, ranging from a lowest value of 0, “Worst imaginable health state”, to a highest value of 100, “Best imaginable health state”. All integer values are indicated by small ticks, while every tenth value (0, 10, 20 ... 100) is labeled and has longer ticks.

The obvious strengths of the VAS, as compared to other valuation methods, lies in the simplicity and flexibility of its mode of administration; unlike, for instance, TTO, the VAS can be understood without much difficulty by respondents in non-interactive self-administrative settings. Visual analogue scales lack support in economic theory and have been criticized by economists for not being choice-based and for being difficult to interpret in relation to the anchors “death” and “perfect health. On the other hand, there is substantial support for the use of VAS, particularly in psychology and psychometrics (there are literally tens of thousands of papers that use different kinds of VAS, often to assess the psychometric properties of other instruments).

VAS has been used in most national valuation studies; in some countries it has been used to help respondents gain familiarity with EQ-5D and health state valuation, before proceeding with primary valuation using TTO. In other countries, it has been used as the primary valuation method.
1.5.3 **Comparisons of Valuation Methods**

Using different valuation methods results in different health state values. Since there is no gold standard, researchers have been interested in how and why the values differ. There is a substantial body of literature investigating differences between the most frequently used valuation methods, VAS, rank, SG, and TTO\(^{16,88-90}\). Such comparisons often include theoretical strength, feasibility, reliability, and internal consistency for each of the methods. Values from the different methods have typically been compared in terms of absolute levels, correlations, and functional form. VAS has usually been considered least burdensome for respondents. SG values typically display the highest numbers of inconsistencies, followed by TTO. Agreement with rank values has been suggested as a criterion for consistency\(^{90}\). A problem with the SG and TTO is that respondents are often unwilling to trade. Generally, SG elicits values that are higher than the values of the other two methods, and SG values usually correlate fairly well with TTO values. Different explanations have been offered for the systematic differences in values elicited with different methods. Some are based on theoretical considerations – the methods measure different constructs, or include different aspects. For instance, TTO is sensitive to time preferences and the SG is sensitive to risk aversion.

1.6 **Construct Relevance and Procedural Invariance**

The 1999 edition of *Standards for Educational and Psychological Tests*\(^{91}\), a joint publication of the American Educational Research Association, the American Psychological Association, and the National Council in Education in Measurement, argues that validity is a unitary concept that does not exist separate from, and therefore cannot be discussed in isolation from, the intended interpretation of a measure. The *Standards for Educational and Psychological Tests* argues that validity should not be subdivided into categories like predictive validity, content
validity, face validity, and criterion validity. Accordingly, criterion, content, and predictive evidence may support specific interpretations of measure scores, but should be considered inseparable from measures’ intended interpretations. Two core concepts are emphasized: construct-irrelevant variance and construct underrepresentation. Construct-irrelevant variance is encountered when the scores on a test or measurement instrument are influenced by any factor not considered relevant to the intended construct. Construct underrepresentation is the failure of a test/measure to capture important facets of the intended construct. Whether or not any specific factor should be considered relevant may be open to debate. In this section, we will present some potential sources of influence that may often be considered construct-irrelevant, and that may illustrate or explain breaches of procedural invariance in health state valuation.

1.6.1 HEURISTICS AND COGNITIVE BIASES

Ordinary life includes a multitude of choices and requires people to make judgments frequently, often without complete information and often under time pressure. Making rational judgments may require considerable effort. Heuristics are efficient, simple rules that are proposed to explain how people manage to make judgments swiftly and with apparent lack of effort. Importantly, these simple rules are beneficial because they usually yield correct results at low cost/effort. However, being simplifications of reality, heuristics may result in systematic errors or cognitive biases.

Cognitive biases may be described as systematic deviations in judgment that occurs in particular situations, resulting in distorted perception, illogical interpretation, and inaccurate judgments. Since health state valuation involves judgments with limited information, limited time, and often complex tasks, heuristics and cognitive biases may threaten valuation methods’ validity. We will briefly present some cognitive biases that are of particular importance to health state valuation.
1.6.1.1 FRAMING EFFECTS

Narrowly defined, framing effects apply when people make different judgments in logically equivalent situations where the only difference is whether something is worded or presented positively or negatively (for instance ‘90% employment’ vs. ‘10% unemployment’)\textsuperscript{99,100}. However, framing effects may also be used to refer to differences in judgment caused by other kinds of differences in wording, or even in all kinds of situation-induced cognitive biases, including for instance anchoring (presented below). The widest definition includes differences in judgments of situations that are equal from the perspective of economic theory\textsuperscript{101}. For example, differences in how people respond to variants of TTO may fall under the category of framing effects. Framing effects have been studied and identified in a variety of fields and settings, including the measurement of patient preferences for treatment methods\textsuperscript{102}. In their presentation of a theoretical framework for TTO\textsuperscript{103}, Buckingham and Devlin point to framing effects as a potential problem when comparing their theoretically (from an economic point of view) equal TTO variants.

1.6.1.2 ANCHORING

Anchoring bias was first described by Kahneman and Tversky, who stated:

*In many situations, people make estimates by starting from an initial value that is adjusted to yield the final answer. The initial value, or starting-point may be suggested by the formulation of the problem, or it may be the result of a partial computation. In either case, adjustments are typically insufficient. That is, different starting-points yield different estimates, which are biased towards the initial values. We call this phenomenon anchoring.*\textsuperscript{93}

Anchoring has been observed in a variety of judgment settings, including some valuation tasks\textsuperscript{104–107}. 
1.6.1.3 FOCUSING

Focusing bias, or focusing illusion, may be defined informally as a tendency to overestimate the importance of the object of our current attention. In the words of Schkade and Kahneman:

\textit{Nothing in life is quite as important as you think it is while you are thinking about it.} \textsuperscript{108}

It has been suggested that focusing contributes to observed differences between general population and patient values for health states: respondents from the general population may focus only on the negative aspects of life in impaired health, whereas patients with impaired health also consider other, non-negative, aspects of life, including things that make them feel good \textsuperscript{109–111}. This is not to say that focusing does not affect patients’ valuations; several studies of focusing have shown that making respondents focus on certain aspects of their lives right before asking about happiness substantially influences how they rate their happiness \textsuperscript{95,108}.

1.6.2 COMMUNICATIVE IMPLICATURES AND THE COOPERATIVE PRINCIPLE

Since the goal of communication is to convey meaning, listeners and readers must try to understand what writers or speakers mean. This usually requires inferences that go beyond the literal meaning of what is said to reach a pragmatic understanding. To successfully draw inferences, communicating parties must, at least to some extent, follow common sets of rules. The most influential description of such rules of communication was expressed by Paul Grice as the \textit{cooperative principle of communication}\textsuperscript{112,113}. The cooperative principle describes pragmatic rules used in natural communication settings, formulated as a set of imperatives on the quantity, quality, relevance and manner of the information that is given. Sperber and Wilson\textsuperscript{114} consider that all the Gricean imperatives are subject to the principle of relevance, which is the most important of the maxims
in the context of this thesis. Adherance to the cooperative principle leads to communicative implicatures: the meaning of utterances that go beyond the literal meaning.

Most research settings, including those involving valuation tasks, deviate substantially from ordinary conversation in several ways. However, respondents in research and survey settings often adhere to, or make assumptions based on, the same principles they use in casual conversation; they assume that information presented is complete, unambiguous, and relevant. Researchers, on the other hand, often do not follow these norms, and very often disregard the possibility that respondents adhere to them. This adherence on the part of respondents and lack of adherence on the part of researchers may lead respondents to answer questions in surprising ways, bias results, or make interpretation of responses difficult.\textsuperscript{115–119} In particular, respondents may interpret wordings, visual presentations, response categories, previous questions or texts, etc. as cues to how they should respond – i.e., “If the experimenter presents it, I should use it.”\textsuperscript{118(p16)} Since, in conversations, the same questions should not be asked more than once, respondents may often interpret repeated questions or similar questions as requests for new information, inducing variance.\textsuperscript{120} In the words of Wänke

Accordingly, respondents will use any cue in the provided information to infer the intended meaning of a question, and to disambiguate vague concepts or ambiguous answering scales. Previous question wording, introductions, the question wording, answer formats, and any other information may serve this purpose.\textsuperscript{119}

This tendency of respondents may be important in valuation tasks: respondents are likely to interpret questions/scales as presenting information on how they should respond. For instance, when repeatedly asked to determine if presented health states are better than, worse than, or equal to death, respondents may easily construe such a task as a request to use all three categories. Otherwise, why present those specific options?
1.6.3 INVOKING DEATH

Within the QALY frame, health states must be measured relative to ‘death’ and ‘perfect health’. However, using ‘death’ as an anchor could be problematic.121 ‘Death’ is not a psychological constant122, and it may be a questionable practice to make all other values dependent on this unstable value. Respondents with extreme views on ‘death’ may distort aggregated health state values123. ‘Death’ is qualitatively very different from living states, since it is impossible to experience the state of “being dead”. People must turn elsewhere for their value judgments of ‘death’. Valuation studies usually give no instructions as to what aspects of ‘death’ respondents should include in their valuations. Some respondents think of it as “simply not being alive”, some include possible distress or grief for their families, and some include religious beliefs about paradise making ‘death’ better than all possible living health states46. VAS or ranking tasks allow analyses of respondents’ value of ‘death’, while the value of ‘death’ is intrinsic to the TTO method and cannot be separated from health state valuations. Since death is an unknown quantity used as an anchor, variance in perceptions of death will be shifted to measures for health states123,124.

The concept of ‘death’ is discomforting for many individuals. For instance, fear of death is common in psychological treatment. A body of literature within the field of psychology indicates that merely invoking death can increase arousal, and thereby have substantial effects on how people think and behave125. Invoking death may have a priming effect that triggers defense mechanisms and leads respondents to change their preferences126,127. Arousal may also have a specific impact on economic behaviour128. Moreover, research suggests that the effects of invoking death increase with the salience of death129–132. The increased influence that occurs with increasing salience may be due to a focusing effect.

In the MVH TTO protocol, each health state valuation starts with an explicit comparison to ‘death’, making the ‘death’ anchor extremely salient. TTO values
could therefore be influenced by unintended and unknown effects related to invoking death. For instance, it has been suggested that discomfort about death may cause the observed ‘gap effect’ in TTO valuation studies. The ‘gap effect’ describes a distributional characteristic of TTO valuation studies, where respondents show aversion to value health states close to zero or death. Since the final TTO value is the result of a series of iterations, the gap effect could indicate that the rest of the comparisons are in some way affected by invoking death at the beginning of the task. Much as individual factors vary when valuing ‘death’ as an anchor state, the effects of invoking ‘death’ could vary across individuals. For instance, individual beliefs, personality, or personal experience are likely to influence individual responses to invocations of death.

We hypothesized that ‘death’, as implemented in the TTO, could cause uncertainty in health state measures. First, variation in what aspects people include when they (implicitly) value the ‘death’ anchor in the TTO task would transfer to all health state values. Second, disparities in how respondents are affected by invoking ‘death’ could add to this uncertainty. Finally, emotional arousal has been shown to impair judgment on tasks requiring numeracy skills, increasing random errors.
2 Objectives

The overarching objective was to enable discussions of some central issues in health state valuation: influence from construct-irrelevant factors, the initial choice task in MVH-style TTO, the validity of WTD values, issues related to valuing death, tariffs as measures of the methodological preferences of the researchers, and the choice of reference group. The included papers were used as background for these discussions.

In this chapter, we present the specific study aims and briefly describe the hypotheses supporting these aims, based on the literature presented in the background section.

1. To investigate how much tariff information is lost when disregarding variance in MVH TTO values within BTD and within WTD.

A cluster of hypotheses made us suspect that discarding within-WTD and within-BTD variance would have limited impact on the relative distance between mean health state estimates. These hypotheses included:

- Using different methods for WTD and BTD states (see section 1.4.2) could confuse respondents, leading to increased noise within BTD and WTD.

- Invoking death in the initial choice task may induce discomfort and arousal, altering rational processes, reducing respondents’ numeracy and ability to perform subsequent valuation (see section 1.6.3 on invoking death) and increasing noise within BTD and WTD.

- Using death as an anchor in valuation shifts variance related to respondents’ perceptions of death to the measures of health states (see section 1.6.3), increasing noise within BTD and WTD.
If respondents adhere to the cooperative principle\textsuperscript{112,113,119} in health state valuation (see section 1.6.2 on communicative implicatures) and assume that the maxim of relevance is in play, the initial choice task is of particular importance (else, why use it for all health states?) and may be construed as a request for respondents to use all three available categories (BTD, ETD, WTD). This would increase the importance of the initial choice task and reduce the importance of subsequent valuation.

- Judgments regarding how much worse than death a health state can be are difficult. Therefore, we would expect considerable noise within WTD.

- The WTD task involves simultaneous manipulation of two of the concepts in play, length of life in the target state and in perfect health (see section 1.4.2 on MVH TTO). In general, people are bad at judgments involving fractions\textsuperscript{67,68,136}. The WTD task involves comparisons of fractions of difficult concepts, increasing the chance of noise within WTD.

2. To investigate the impact of different arbitrary transformation methods for health states considered to be WTD, as compared to different regression models and different population.

The Torrance transformation used in the US valuation study results in higher TTO values than the Patrick transformation used in the UK valuation study\textsuperscript{69} (section 1.4.2.1). The US tariff has higher values than the UK tariff\textsuperscript{137,138}. We hypothesized that the transformation method would contribute substantially to observed tariff differences because of the large number of WTD valuations.

3. To investigate learning effects in TTO valuation – do respondents change the way they respond as a function of how many health states they have valued previously?
Since the MVH TTO protocol uses different methods for valuing BTD and WTD health states (section 1.4.2), learning effects could differ between BTD and WTD.

4. To explore ranking and TTO values from previous valuation studies to assess whether the differences between the two methods make respondents sensitive to different EQ-5D dimensions

TTO focuses greatly on length of life and comparison to death (see section 1.4.2). Ranking focuses solely on the descriptive attributes of presented health states (section 1.5.1). We hypothesized that this difference could lead respondents to focus on different dimensions of the EQ-5D, using the two valuation tasks.

5. To compare patterns in the relative importance attributed to the different EQ-5D dimensions by the general public valuing hypothetical health states and by respondents describing their own health.

Previous comparisons of patient and general population values have focused on absolute values attributed to the different dimensions or on the functional form of relationships between values elicited from the two groups of respondents (section 1.3.2.1). We hypothesized that a different picture could emerge if focusing on what dimensions of health were considered most and least important for quality of life by the two groups of respondents.

6. To investigate the relationship between attitudes toward euthanasia and EQ-5D values elicited using MVH TTO, LT-TTO, and VAS.

Comparison to death is direct in MVH TTO, indirect in LT-TTO, and performed separately in VAS. We therefore hypothesized that attitudes toward euthanasia would influence these three valuation tasks differently.

7. Investigating anchoring effects in LT-TTO and MVH TTO.

Anchoring (section 1.6.1.2) has been observed in numerous settings,
including some valuation tasks⁹³,⁹₂,¹⁰⁴,¹⁰⁶. Since the initial offer of time in perfect health has not been varied in TTO-based EQ-5D valuation studies, potential anchoring has not been discernible. We hypothesized that altering the initial offer would reveal anchoring effects in TTO valuation.
3 METHODS

This chapter describes the external datasets used in five of our papers, our valuation experiments used in the remaining two papers, and the methods of analyses used in each paper. For reference, we include two tables at the beginning of the chapter: table 2, listing characteristics of the datasets used and in what sections they are described, and table 3 listing papers, datasets, and valuation methods.

Table 2 – Data sources

<table>
<thead>
<tr>
<th>Study</th>
<th>Method</th>
<th>n*</th>
<th>EQ-5D states</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>External datasets</td>
<td>Rank/VAS/MVH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK valuation</td>
<td>TTO</td>
<td>2,997</td>
<td>12 Hypothetical</td>
<td>3.1.1</td>
</tr>
<tr>
<td></td>
<td>Rank/VAS/MVH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US valuation</td>
<td>TTO</td>
<td>3,773</td>
<td>12 Hypothetical</td>
<td>3.1.2</td>
</tr>
<tr>
<td>MEPS</td>
<td>VAS</td>
<td>24,277</td>
<td>Experienced health</td>
<td>3.1.3</td>
</tr>
<tr>
<td>Valuation experiments</td>
<td>Rank/VAS/MVH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MVH TTO group</td>
<td>TTO</td>
<td>218</td>
<td>8 Hypothetical</td>
<td>3.2.3.1</td>
</tr>
<tr>
<td></td>
<td>Rank/VAS/LT-TTO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LT-TTO group#</td>
<td>TTO</td>
<td>411</td>
<td>8 Hypothetical</td>
<td>3.2.3.2</td>
</tr>
<tr>
<td>TTO5 group</td>
<td>Rank/VAS/TTO</td>
<td>396</td>
<td>8 Hypothetical</td>
<td>3.2.3.3</td>
</tr>
</tbody>
</table>

* After exclusions
# Subdivided into 11 groups with different starting point
### Table 3 - Papers, datasets, and valuation methods

<table>
<thead>
<tr>
<th>Paper and topic</th>
<th>Datasets</th>
<th>Valuation methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Collapsed TTO</td>
<td>US valuation study</td>
<td>MVH TTO</td>
</tr>
<tr>
<td></td>
<td>UK valuation study</td>
<td></td>
</tr>
<tr>
<td>II WTD transformation methods</td>
<td>US valuation study</td>
<td>MVH TTO</td>
</tr>
<tr>
<td></td>
<td>UK valuation study</td>
<td></td>
</tr>
<tr>
<td>III Learning effects in TTO</td>
<td>US valuation study</td>
<td>MVH TTO</td>
</tr>
<tr>
<td>IV Rank vs. TTO</td>
<td>UK valuation study</td>
<td>Ranking/MVH TTO</td>
</tr>
<tr>
<td></td>
<td>US valuation study</td>
<td>Ranking/MVH TTO</td>
</tr>
<tr>
<td>V Hypothetical vs. Experienced health</td>
<td>US valuation study</td>
<td>VAS (hypothetical)</td>
</tr>
<tr>
<td></td>
<td>MEPS</td>
<td>VAS (experience)</td>
</tr>
<tr>
<td>VI Euthanasia and valuation</td>
<td>LT-TTO group</td>
<td>VAS/LT-TTO</td>
</tr>
<tr>
<td></td>
<td>MVH TTO group</td>
<td>VAS/MVH TTO</td>
</tr>
<tr>
<td>VII Anchoring in LT-TTO and MVH TTO</td>
<td>LT-TTO group*</td>
<td>LT-TTO</td>
</tr>
<tr>
<td></td>
<td>MVH TTO group</td>
<td>MVH TTO</td>
</tr>
<tr>
<td></td>
<td>TTO-5 group</td>
<td>TTO-5</td>
</tr>
</tbody>
</table>

### 3.1 EXTERNAL DATASETS

Many of our analyses were performed on data from previously conducted studies. In particular, we used data from the US and UK EQ-5D valuation studies extensively. We will describe the UK valuation study in some detail, since it is the model on which most subsequent valuation studies, including the US one, have been built. In our description of these datasets, we will focus on aspects that are important to our objectives and studies. Therefore, several other methodological choices that are important in their own right, for instance, the advanced sampling techniques aimed at achieving population representativeness, will receive limited attention.
3.1.1 UK EQ-5D VALUATION STUDY

This dataset was used in paper I, II, and IV. The goal of the UK Measuring and Valuing Health (MVH) EQ-5D valuation study was to model the mean preferences for all EQ-5D health states of the non-institutionalized adult population of England, Scotland, and Wales. Cost considerations, combined with pilot studies indicating that respondents could not be expected to value more than 13 health states each, led the researchers to conclude that measuring mean preferences for all 243 EQ-5D health states would be virtually impossible. Therefore, they selected 42 EQ-5D health states for direct valuation and devised regression models to interpolate the remaining values.

Using a representative sample of the population, a sample size of 3,235 was needed to achieve an 80% chance of detecting a 0.05 difference in mean values between health states. Between August and December 1993, a total of 3,395 respondents were interviewed in their own homes by 92 trained interviewers. Complete data were available for 2,997 respondents, and these were used in all subsequent analyses.

As a warm-up exercise, each respondent was first asked to describe his/her own health using the EQ-5D descriptive system. Then, they were presented with cards describing 15 “health states”, including state 11111 (perfect health), “immediate death”, “unconscious”, and 12 EQ-5D health states selected from the total pool of 42. They were instructed to imagine being in each of these states for 10 years without change, followed by death. First, they were asked to rank these cards from best to worst. Second, with the cards still in their ranked order, they were asked to rate each on the VAS. The 12 EQ-5D health states plus “unconscious” were then valued in randomized order, using the MVH TTO protocol. Prior to performing analyses, negative TTO-values (values for states considered WTD) were transformed using the Patrick transformation\textsuperscript{65}. 
With this set of data, consisting of 2,997*12 (transformed) TTO valuations, the MVH researchers generated an algorithm for calculating values for all 243 EQ-5D health states. Since each respondent valued 12 health states, the different values given by each respondent were considered not completely independent. Ordinary least-squares (OLS) regression requires independence between measures. The authors therefore opted for an additive GLS (generalized least-squares) regression, with random effects specified at the level of individual respondents. After testing several models, they decided on the following: a constant term, two dummies per dimension, one of which reflected the dimension being at level 2 (some problems), and the other reflecting the dimension being at level 3 (extreme problems), and N3, reflecting the presence of any dimensions at level 3. The model was used to predict values for all health states except 11111 (no problems on any dimension), which was given a defined value of 1. The constant term was interpreted as the mean loss of HRQoL associated with any movement away from perfect health.

3.1.2 US EQ-5D VALUATION STUDY

This dataset was used in papers I, II, III, IV and V. The US EQ-5D valuation study, which was performed in 2002, was based on the UK study protocol. We will point to differences between the US and the UK studies. A complex clustered sampling procedure was used, with oversampling of minority groups, and individual respondents were weighed to ensure population representativeness. Interviews were conducted both in English and Spanish. A total of 4,048 respondents went through the valuation procedure, 3,773 of whom were included in the tariff calculation. Missing data was imputed using a hot deck procedure. For details on sampling, exclusions, and imputation, we refer to the original study paper\textsuperscript{51}.  

62
WTD values were transformed using the Torrance transformation\textsuperscript{64}, instead of the Patrick transformation\textsuperscript{65}, which was used in the UK study. Details can be found in section 1.4.2.1.

A different regression model specification was used to model mean values for the EQ-5D health states. The US D1 model used the same 10 main dimension dummies as the UK N3 model but replaced the N3 term and the constant term with four new variables, called I22, I3, I32, and D1. I2 (a variable that was dropped due to lack of statistical significance) represented the number of dimensions, beyond the first, at level 2, and I3 represented the same at level 3. I22 and I32 were I2 and I3 squared. D1 represented the number of dimensions not at level 1, beyond the first. A GLS regression procedure was used, with procedures to handle the complex sampling and respondent weights. It can be shown that the D1 variable equals the negative of the constant term, a process that shifts all 10 main dimension dummies away from 0 by a value equal to the constant. We have criticized this elsewhere\textsuperscript{139}, because it makes at least two of the dummies pass the test of statistical significance, increases the multicollinearity of the model considerably, and makes the coefficient harder to interpret.

Participants in the US study valued the same 42 EQ-5D health states as in the MVH study and the same number of health states per individual (12 + 11111, “death”, and “unconscious”). However, the selection of health states was not randomized at the individual level. Instead, respondents were allocated to 5 groups, each of which valued the same subset of health states in an individually randomized order.

### 3.1.3 Medical Expenditure Survey Panel

The medical expenditure survey panel (MEPS) is a set of large-scale surveys of nationally representative samples of the US general population, doctors,
hospitals, employers, and other health services. The objective of MEPS is to allow analyses of usage, availability, and cost of health care services and insurance systems in the US. MEPS is divided into two main parts, an insurance component and a household component. We were interested only in the household component, which contains data on the health, demographics, and health care usage of individuals and their families. Information on MEPS and most of the collected data is available on the MEPS public website (http://meps.ahr.gov).

Over a period of four years, from 2000 to 2003, the MEPS surveys included collection of self-ratings on EQ-5D and the EQ-VAS. A total of 74,277 complete EQ-5D and EQ-VAS ratings are available for analyses, with respondents reporting 196 different EQ-5D health states. We used this set of data in paper V, as a source of self-rated HRQoL.

3.2 VALUATION EXPERIMENTS

To shed light on a number of issues related to TTO and other methods used to value EQ-5D health states, we performed several valuation experiments/surveys. Only two of the issues we wanted to look into are handled in papers in this thesis: how attitudes toward euthanasia may influence valuation (paper VI) and investigation of anchoring in TTO and LT-TTO (paper VII). Many of our choices when conducting these studies were influenced by financial constraints. In this presentation, we will focus primarily on the choices and methods used in the parts of the set of surveys that have been used in papers in this thesis. The parts of the surveys that we have not yet looked into will be presented very briefly where relevant.

3.2.1 SAMPLING OF RESPONDENTS

We hired the Norwegian branch of the international survey institute Synovate to administer the surveys. Members of Synovate’s standing pool of willing survey respondents were invited to participate by email. For each survey (details to
follow in a later section), Synovate was asked to invite a sample that roughly mimicked the makeup of the adult Norwegian population, with regard to age, sex, and geography, and with intentional oversampling of the youngest and the oldest groups because these have been known to be less likely to participate. As incentive to participate, three respondents were selected by ballot to receive universal gift-cards (one worth NOK 10,000, and two worth NOK 5,000).

3.2.2 STRUCTURE OF THE SURVEYS
The interviews used in our valuation experiments were structured much like the US and UK TTO-based valuation studies and were divided into three parts. The first part of each interview focused on demographics and background variables, the second part contained the valuation tasks, and the third looked at other measures of interest to us, specifically the personality trait neuroticism (not used in this thesis), attitudes toward euthanasia, and religious outlooks (not used in this thesis).

3.2.2.1 DEMOGRAPHICS
The initial part of each interview and covered demographics and other background variables, including age, sex, education, vocational status, marital status, number and ages of children, birth nationality, geographic region, smoking status, and a few questions about experiences with reduced health. The appendix contains the list of questions asked in the initial part of the survey (in Norwegian).

3.2.2.2 VALUATION
The valuation part of the survey started by asking respondents to describe their own health, using the EQ-5D questionnaire. Then, respondents were presented with the eight target EQ-5D health states, as well as state 11111 (perfect health) and “immediate death”, and were asked to rank these from subjective best to worst. Subsequently, they were asked to rate all 10 ranked states on a VAS: the
top ranked first, then the worst ranked, then they were asked to select a state that was perceived as roughly half-way between the two, and so on.

After ranking and VAS-rating, TTO valuation of the eight target states ensued. The eight health states were presented in individually randomized order. Several different TTO variations were used on separate groups of respondents. The groups and TTO variants are described later.

3.2.2.3 ATTITUDES TOWARD EUTHANASIA

The third part of the survey was divided in three subsections, measuring neuroticism, religious outlooks, and attitudes toward euthanasia. This thesis does not address issues related to the measures of neuroticism or religious outlooks. We will present these briefly because they preceded the measurement of attitudes toward euthanasia.

All three sections consisted of statements with which the respondents were asked to state their level of agreement. For each statement, there were five options: “strongly agree”, “agree”, “neither agree nor disagree”, “disagree”, and “strongly disagree”.

Neuroticism was assessed using the NEO Five Factor Inventory (12 questions), which is a short-form personality measure based on the five-factor personality model. Ideally, we would have included all five personality factors, but due to cost considerations, we decided to include neuroticism only to reduce the total length of the interview. There were two statements about religion: “I consider myself to be personally religious” and “I believe in life after death”. Analyses of neuroticism and the questions about religion are not included in this thesis.

Finally, respondents were presented with a case about a dying person in great pain and they were asked three questions regarding the extent to which they agreed with statements that doctors should be allowed to perform passive euthanasia, active euthanasia, or assisted suicide.
3.2.2.4 Selection of health states for valuation

Unlike typical valuation studies, the objectives of this study were related to the valuation methods and not to the EQ-5D per se. Whereas valuation studies with the objective of creating tariffs for all EQ-5D health states require valuation of as many EQ-5D health states as possible, this was not a requirement of our studies. We selected eight EQ-5D health states from the pool of 42 states that have been used in several previous valuation studies. In the selection of health states, we used three criteria: First, we wanted to sample a wide range of severity. Second, since several of our hypotheses were related to either the initial choice task in MVH TTO (better/worse than death) or to valuation of WTD health states, we wanted several states that many respondents would seriously consider rating as worse than death. Third, we wanted a sample of health states where all five dimensions were represented to some degree. Three states were selected based on other criteria: the worst possible, “pits”, EQ-5D health state, 33333; the “middle” state, health state 22222, and the state that was considered as next best to ‘perfect health’ in the US and UK valuation studies, state 11211. For the remaining five states, we used lists of mean values for the 42 health states valued in the US and UK valuation studies as our starting point, and selected states with a relatively wide spread of mean values, with oversampling of the more serious states. Table 4 lists the eight EQ-5D health states, along with their mean TTO values from the US and UK valuation studies.
Table 4 – Valuation experiment EQ-5D health states

<table>
<thead>
<tr>
<th>Health state</th>
<th>UK mean value</th>
<th>US mean value</th>
</tr>
</thead>
<tbody>
<tr>
<td>11211</td>
<td>.87</td>
<td>.87</td>
</tr>
<tr>
<td>11312</td>
<td>.55</td>
<td>.65</td>
</tr>
<tr>
<td>22222</td>
<td>.50</td>
<td>.60</td>
</tr>
<tr>
<td>32211</td>
<td>.15</td>
<td>.33</td>
</tr>
<tr>
<td>23232</td>
<td>-.09</td>
<td>.22</td>
</tr>
<tr>
<td>32223</td>
<td>-.18</td>
<td>.20</td>
</tr>
<tr>
<td>33333</td>
<td>-.54</td>
<td>-.10</td>
</tr>
</tbody>
</table>

3.2.3 TTO VARIANTS

We had separate groups of respondents perform TTO valuation using three different TTO types, referred to as MVH TTO, LT-TTO, and TTO5.

3.2.3.1 MVH TTO

Valuation with MVH TTO was modeled directly on the protocol used in the MVH study as presented in the background chapter. However, unlike most valuation studies, our experiments were performed using a web-applet, and respondents answered the questions in their own homes.

3.2.3.2 LT-TTO

As mentioned in the background section, LT-TTO is new, and several details are still subject to experimentation. We chose to use a life B that consisted of 10 years lead-time, followed by 10 years in the target state. To test for possible anchoring effects, we randomized respondents to 11 different starting points for the length of life A, from a minimum of 10 years to a maximum of 20 years. Thus, some respondents were initially presented with a life A of 10 years in perfect health compared to life B (10 years perfect health, 10 years target state), while others were presented with an initial life A of 11 years, 12 years ... 20 years in perfect health. As in the BTD task of the MVH protocol, we used an incremental
routing, with 1 year increments to determine the point of preferential equilibrium, with a correction of ½ year if required.

While we hypothesized that the length of available lead-time would influence elicited values, we did not experiment with different lead-time lengths, since this would require a greater number of respondents than we had available.

3.2.3.3 TTO5

In addition to MVH-style TTO and our 11 LT-TTO variants, we used an altered version of MVH-TTO that we will refer to as ‘TTO5’. One of our primary hypotheses was that the initial choice task in the MVH protocol could create problems by the explicitly invoking “immediate death”. If the starting point and non-linear routing procedure used in the MVH protocol created anchoring effects, these effects would be difficult to disentangle. In our LT-TTO experiments, we varied only the starting point, hoping to capture potential anchoring. In TTO5, we shifted the starting point for life A from death to 5 years in perfect health. From this starting point, life A was altered with 1-year increments. If the respondent traded away all life-time in life A and stated a preference for immediate death to 10 years in the target state, the MVH WTD protocol was used, starting with life A at 9 years in the target state followed by 1 year in perfect health.

3.2.4 RESPONDENT GROUPS

Synovate was asked to collect at least 300 respondents to MVH-style TTO valuation, 300 respondents using TTO5, and at least 400 respondents using LT-TTO. We wanted more LT-TTO respondents because the LT-TTO group was divided into 11 subgroups. Total sample size was determined by financial constraints, meaning that we were free to distribute respondents between groups, as long as the total number of respondents was unaltered. Lacking theoretical or empirical basis for power analyses, the distribution of respondents
to the different groups was determined by the rather unscientific method of educated guessing\textsuperscript{143,144}.

3.2.5 EXCLUSION CRITERIA

We used a uniform set of exclusion criteria in our valuation experiments, in order to avoid data from respondents who clearly failed to perform the tasks as intended. Respondents were excluded if they triggered one or more of these three criteria:

1. All eight health states were assigned values equal to or worse than death (\(\leq 0\)).
2. All eight health states were assigned the same value.
3. The best state (state 11211) was assigned a value equal to or worse than the worst (state 33333).

Criteria 1 and 2 are similar criteria used in most valuation studies based on the MVH protocol. Criterion 3 was included because failure to distinguish between these two health states was considered an indicator of respondents misunderstanding the task or responding at random.

3.3 ANALYSES BY PAPER

3.3.1 PAPER I: COLLAPSED TTO

The objective of paper I was to assess the information loss incurred by disregarding variance within BTD and within WTD in TTO-based EQ-5D valuation. Analyses were performed on individual TTO valuation data from the US and UK EQ-5D valuation studies. For each health state valuation (12 total per respondent), hereby referred to as ‘fTTO’ value (full TTO value), we calculated corresponding collapsed TTO values (hereafter ‘cTTO’):
cTTO thus disregards variation within WTD and within BTD. For each dataset, we calculated means of fTTO and means of cTTO for all 42 valued EQ-5D health states. We also calculated regression-based tariffs using the regression model specifications from the US and UK valuation studies, respectively, based on fTTO and cTTO values. We then used regression modeling to predict fTTO means from cTTO means, and fTTO tariffs from cTTO tariffs.

Additional analyses were performed to model a scenario in which a large group of respondents elicited cTTO values, and a small set of respondents elicited fTTO values, to crudely illustrate a possible implementation of the cTTO approach.

Finally, we calculated fTTO means of BTD values and of WTD values for each of the 42 measured health states to assess the sensitivity of these procedures to different severity of health states.

3.3.2 PAPER II: WTD TRANSFORMATION

The objective of paper II was to compare the relative contributions of national dataset, regression model specification, and WTD transformation method to the observed differences between the US and UK EQ-5D tariffs. Analyses were performed on individual TTO valuation data from the US and UK EQ-5D valuation studies, in which different WTD transformation methods and different regression models were used.

We calculated tariff values for all 243 EQ-5D health states using all eight possible combinations of national dataset, regression model specification, and WTD transformation method. Using these eight tariffs, we performed three sets of comparisons of contribution to observed differences:
1. Graphical comparison of the eight tariffs.

2. For each of the three sources of difference, we compared the tariffs with the variant from the US study with the tariffs with the variant from the UK study. For instance, the four tariffs with US data were compared with the four UK data tariffs.

3. We calculated the mean distance between all possible (n=29,403) movements from one EQ-5D health state to another within each tariff. We then calculated mean differences between the mean distance values from the US variants with the mean distances from the UK variants for each of the three sources of difference.

3.3.3 **Paper III: Learning Effects in TTO**

The objective of paper III was to determine if respondents in the US EQ-5D valuation study changed their TTO responses as a function of how many health states they had previously valued (the sequence number). We used individual TTO valuation data from the US valuation study.

To test if TTO values changed as a function of sequence, we used multiple linear regression with TTO values as the dependent variable, and the sequence number as the independent variable. We then performed the same analyses split for only BTD values and only WTD values.

To test for possible fatigue effects (or if respondents learned “short-cuts”), we used linear regression predicting the number of iterations (TTO choice tasks) necessary to reach the elicited TTO values using the sequence number as the predictor.

3.3.4 **Paper IV: Ranking vs. TTO**

The objective of paper IV was to explore ranking and TTO values from previous valuation to determine if different methods lead respondents to become more
sensitive to impairments of different EQ-5D dimensions. We used individual TTO and rank data from the US and UK EQ-5D valuation studies.

Rank and TTO data are not directly comparable. We assigned the 42 valued health states in each dataset overall rank (called “mean rankings”) and overall TTO rank (“mean TTO rank”) values: using the mean TTO and mean rank values for the 42 states, the best health state by each method was assigned an overall rank value of 1, and the worst an overall rank value of 42. We then subtracted mean rankings from the mean TTO ranks. A positive rank difference would thus signify a health state that was, on average, ranked as worse using TTO than using the ranking task.

With the lists of mean TTO ranks, mean rankings, and rank differences, we performed two operations:

1. We identified all pairs of health states for which the mean TTO ranks and mean rankings were reversed to see if they had any discernible pattern with regard to what dimensions of health were impaired.
2. We performed multiple linear regression modeling predicting rank differences using 10 dummy variables representing impairments on level 2 (some problems) and level 3 (extreme problems) for the five EQ-5D health dimensions as predictors.

3.3.5 **Paper V: Experienced and Hypothetical Health**

The objective of paper V was to compare the patterns of relative importance attributed to the different EQ-5D dimensions by the general public (hypothetical health, HH) and by respondents describing their own health (experienced health, EH).

We used VAS self-rated health with corresponding EQ-5D health descriptions from the Medical Expenditure Survey Panel (MEPS) as our EH data, and VAS-
ratings of hypothetical EQ-5D health states from the US EQ-5D valuation study as our HH data.

For EH and HH data, we performed multiple regressions using VAS values as our dependent variable, and independent variables from a mathematical equivalent of the D1 regression model specification from the US valuation study. The regression model included ten dummy variables representing the five EQ-5D dimensions at level 2 (some problems) and level 3 (extreme problems), the variables \( I_2, I_3, I_{22} \) and \( I_{32} \). \( I_2 \) and \( I_3 \) represented the number of dimensions at level 2 and 3, respectively, beyond the first. \( I_{22} \) and \( I_{32} \) were \( I_2 \) and \( I_3 \) squared.

The D1 variable was dropped in favor of a constant representing any movement away from perfect health.

We then compared the ten main dimension dummy variable coefficients from our EH model with the corresponding coefficients from the HH model. To enable comparisons of relative dimension importance, we divided the five level 2 coefficients by their mean, and the five level 3 coefficients by their mean, resulting in adjusted coefficient values with an average of 1. As a measure of relative dimension importance, we used the mean of the adjusted level 2 and adjusted level 3 coefficients, separately for the EH and HH models. The EH and HH patterns of relative dimension importance were then compared.

### 3.3.6 Paper VI: Attitudes Toward Euthanasia

The objective of paper VI was to investigate the relationship between attitudes toward euthanasia (ATE) and EQ-5D values elicited using MVH TTO, LT-TTO, and VAS.

We used data collected in two valuation experiments in Norwegian general population samples. These are the MVH-TTO and LT-TTO groups described in detail under section 3.2. VAS data from both groups were pooled. ATE was assessed by asking respondents to state their level of agreement on a five-point
scale with three statements regarding active euthanasia, passive euthanasia, and assisted suicide. Our ATE scale was calculated by assigning the five scale alternatives values from -2 to 2, and averaging across the three statements. Separately for MVH TTO (Patrick transformation), LT-TTO, and VAS, we used multiple regression predicting health state values with ATE, age (years), sex, and education (<11 years, 11-13 years, >13 years, represented by two dummies) as predictors.

3.3.7 PAPER VII: ANCHORING BIAS IN TTO

The objective of paper VII was to investigate potential anchoring effects in LT-TTO and MVH TTO. We used data from three valuation experiments performed in Norwegian general population surveys, described in detail in section 3.2.

Primary analyses were performed on the LT-TTO dataset, which was subdivided into 11 groups, with different starting points, ranging from initial TTO values of 0 to 1 by .1 increments. We used robust regression to predict TTO values using age (years), sex, education (<11 years, 11-13 years, >13 years), and the TTO value of the starting point as predictors.

Secondary analyses were performed by comparing TTO values from our MVH TTO group with the TTO5-group (TTO5 being MVH-TTO with an altered starting point of 5 years of perfect health, corresponding to a TTO-value of .5). TTO values from the two groups of respondents were compared using independent-samples t-test.

3.4 PERMISSIONS AND ETHICS

MEPS is available to researchers at http://www.meps.ahrq.gov/

US valuation study data is publicly available on the Internet at http://www.ahrq.gov/rice/EQ5Dproj.htm
UK valuation study data is available for research through the UK Economic and Social Data Service http://www.esds.ac.uk/Lucene/Search.aspx

The studies on which papers I through V were based have not been considered by any ethical committee, since they were performed on publicly available, anonymized datasets.

The valuation experiments presented in papers VI and VII were performed under the umbrella of the market research institute Synovate, which is licensed by the Norwegian Data Inspectorate to perform surveys in the Norwegian population. All data were anonymized by Synovate before we had access to them. Prior to administration of the surveys, we had telephonic contact with the administration at the Regional Medical Ethics Committee, who determined that this was not medical research falling under their jurisdiction. We then contacted the Norwegian Data Inspectorate directly, to ascertain whether the surveys could be performed by Synovate without further consideration from them or another ethics committee. All invited participants were informed that participation was voluntary, that all gathered data would be anonymized prior to delivery by Synovate, and that it was our intention to make the data publicly available on the Internet.

Since no ethical research committee considered the study to be inside their jurisdiction, we had to perform the ethical considerations. According to Norwegian law, personally identifiable information regarding health is strictly regulated. However, we ascertained that identification was impossible by having Synovate anonymize the data at the first possible junction. Thus, the primary ethical concern was the potential harm or discomfort for participants due to the nature of the questions posed. Our surveys involved TTO, which may discomfort respondents with its direct comparisons of impaired health to “immediate death”. We also asked questions regarding attitudes toward euthanasia and perspectives on the possibility of life after death, both of which may distress
some respondents. To minimize discomfort, we included introductory text specifically stating that the questions were posed for research purposes only, that participation was not in any way mandatory, and we included a choice labeled “do not wish to respond” to the questions on euthanasia and religious outlooks. Several potentially interesting topics and questions were removed from our surveys, since we did not consider the potential gain in knowledge sufficient to justify their potential burden on the respondents. In their final form, we considered the potential burden of our surveys sufficiently small that they could be justified by the need for further knowledge regarding health state valuation.
4 MAIN RESULTS

4.1.1 PAPER I: COLLAPSED TTO

Analyses were performed on data from the US and UK valuation studies. Individual TTO values (fTTO) were collapsed onto four points (cTTO), such that WTD = .5, ETD = 0, BTD = .5, and UTT = 1 (Section 3.3.1). Pearson’s correlation between mean values for the 42 measured health states based on fTTO and the corresponding mean values based on cTTO was >.998 (p<.0001) in both the US and UK datasets, with both Patrick and Torrance transformations of WTD values (Figure 6a). Pearson’s correlation between estimated tariff values for all EQ-5D health states (except state 11111) based on fTTO and corresponding values based on cTTO was >.998 (p<.0001) using both US and UK data with both Patrick and Torrance transformations of WTD values (figure 6b). Mean WTD fTTO values for the 42 health states were similar for the best and the worst health states in both national datasets, using both WTD transformations (figure 7).

Figure 6 – cTTO and fTTO values

<table>
<thead>
<tr>
<th>a)</th>
<th>b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a)</td>
<td>b)</td>
</tr>
<tr>
<td>UK Mean values from fTTO and cTTO</td>
<td>Predicted values from fTTO and cTTO</td>
</tr>
</tbody>
</table>
Figure 7 – Distribution of UK TTO values

Health states ordered by mean TTO value. Circle area proportional to the prevalence of TTO value for health state. Lines represent mean values for presented values.
4.1.2 PAPER II: WTD TRANSFORMATION

Analyses were performed on the US and UK valuation datasets. We calculated 8 tariffs, combining US and UK data, WTD transformation methods, and regression models (section 3.3.2.) Figure 8 illustrates the modeled tariffs. Most of the observed difference between the UK and US EQ-5D tariffs was attributable to the use of different transformation methods for WTD health states. As expected, the differences in tariff values caused by using different transformation methods increased with increasing severity of the health states, since more respondents valued them as WTD.

The mean absolute difference in values was 1.5 times larger when changing transformation methods than when using different datasets. Choice of transformation method had an effect on the mean health gain that was 3.2 times larger than the effect of switching dataset. The mean health gain in the UK tariff was .09 higher than in the US tariff. Using the Patrick (UK) transformation on the US dataset reduced this absolute difference to .02. Choice of regression model had little overall impact on the differences between the value sets.
Figure 8 – Tariffs by dataset, WTD transformation, and regression model
4.1.3 **Paper III: Learning Effects in TTO**

Analyses were performed on US valuation data. We arranged individual TTO values by their sequence number (the place in the sequence of the 13 valued states for each respondent, Section 3.3.3). Mean TTO values were fairly stable across the sequence number, but analyzing BTD and WTD separately revealed that TTO values became more extreme with increasing experience: BTD values increased by .02, while WTD values declined by .21 (p<0.0001) over the full sequence of 13 valuations (Figure 9). Since there were relatively fewer WTD responses than BTD responses, these two separate effects were cancelled out when analyzing the the full mean (BTD and WTD together). Standard deviations increased slightly: OLS regression indicated an increase from .47 to .52 over the full sequence. The number of choice iterations had no statistical association with the sequence number. Findings were stable across levels of health state severity, age and sex.

**Figure 9 – TTO values by sequence number**

Circle area proportional to prevalence of TTO value by sequence number. Dashed line represents mean of BTD values. Solid line represents mean of WTD values.
4.1.4 Paper IV: Ranking vs. TTO

Analyses were performed on US and UK valuation data, using the rank ordering of mean TTO values (mean TTO ranks) and mean rank values (mean rankings) (Section 3.3.4). In total, 43 (US) and 41 (UK) pairs of health states displayed reverse ordering of mean TTO ranks and mean rankings. In most of these pairs, the states ranked as best on TTO displayed impairments on the first three EQ-5D dimensions (mobility, self-care, and usual activities), and the states ranked as best in the ranking task displayed impairments on the last two (pain/discomfort and anxiety/depression). Regression models predicting mean rank differences (mean TTO ranks – mean rankings) consistently indicated relatively better TTO rankings for states impaired on the three first dimensions, and relatively better rank values for states impaired on the last two.

4.1.5 Paper V: Experienced and Hypothetical Health

We used hypothetical health (HH) VAS data from the US valuation study and experienced health (EH) values from the MEPS (Section 3.3.5). Our regression models indicated that EH ratings were higher (smaller losses) than corresponding HH ratings (figure 10). When we divided each main dimension dummy variable coefficient by the mean of all five dummy variable coefficients at the same level (some problems/extreme problems) from the same reference group, the patterns of relative dimension importance were reversed for the two sources (figure 11).
4.1.6 PAPER VI: ATTITUDES TOWARD EUTHANASIA

Analyses were performed on data from the LT-TTO and MVH-style TTO groups in our valuation experiment surveys. We predicted TTO and VAS values by age, sex, education, and scores on our attitudes toward euthanasia (ATE) scale. (section...
After applying exclusion criteria, 411 LT-TTO respondents, 218 TTO respondents and 603 VAS respondents were available for analyses. Relatively more respondents with lower education than high education were excluded. TTO and LT-TTO were associated with attitudes toward euthanasia (ATE), in addition to preferences for health states. A one-point increase on the ATE scale predicted a change in mean TTO value of -.113 (p<0.001, see figure 12) and in LT-TTO of -.072 (p<0.001). The influence of ATE was larger with an increasing severity of the health states. Demographic variables, but not ATE, predicted VAS values. ATE scores were associated with increasing number of WTD valuations per individual for all valuation tasks. The association was strongest for TTO values and weakest for VAS values.

**Figure 12 - TTO by attitudes toward euthanasia**
4.1.7 **PAPER VII: ANCHORING BIAS IN TTO**

Analyses were performed on data from the LT-TTO, MVH-style TTO, and TTO5 groups in our valuation experiment surveys. We predicted TTO values using age, sex, education, and TTO starting point value (section 3.3.7). In our LT-TTO group, 73 of 484 respondents (15.1%) triggered one or more of the exclusion criteria, leaving 411 respondents for the primary analyses. In our MVH TTO group, 110 of 328 were excluded (33.5%), and 41 of 437 (9.4%) were excluded in the TTO5 group. Respondents with starting points closer to 0 were much more likely to trigger all three criteria than respondents starting further up on the scale.

A robust regression model predicting LT-TTO values by age, sex, education, and starting point TTO value identified starting point value as the only statistically significant predictor. An increase of 1 in the starting point value (the span of our starting point value groups) resulted in a mean increase in TTO values of .250 \((p<.001)\). Regression performed without exclusions indicated an increase in elicited TTO values of .374 \((p<.001)\) across the measured starting points. Figure 13 illustrates the distribution of values, starting points, and means of the different groups of respondents. The independent-sample t-tests comparing the MVH TTO to TTO5 groups identified a mean difference in values of .085 \((p<.001)\) after exclusions and of .395 \((p<.001)\) when all respondents were included.
Figure 13 - TTO values by starting point group

Circle area proportional to TTO value by group. Solid line represents mean TTO value. X represents starting point for the TTO task.
5 DISCUSSION

This chapter is divided in two main parts: discussion of methods, and discussion of results. The method discussion starts with considerations regarding topics that are of importance to several of the papers, followed by discussions of methods specific to each paper. The structure of the main discussion will be described in greater detail in the beginning of the subchapter.

5.1 DISCUSSION OF METHODS

5.1.1 EXTERNAL DATASETS (PAPERS I-V)

We had no option but to accept the external datasets “as is”; we had no influence over the selections of respondents or methods or over how the studies were recorded. For instance, in our study of learning effects in TTO valuation (paper III), we were limited to use of the US dataset because our copy of the UK dataset was organized in a manner that did not reveal the order in which the health states were presented. Later, we were informed that raw data including this information was available through the UK Economic and Social Data Service.

When performing analyses of previous valuation study data, we had to decide which set of valuation studies to use. We selected the UK and US studies for several reasons: the UK study is the prototype on which later TTO-based valuation studies have been built. The US and UK datasets were the largest we have access to. The two studies were performed in different populations nearly ten years apart, and they are the valuation studies with which we believe the greatest number of potential readers are familiar. However, the differences in methods used posed a challenge, particularly the use of different WTD transformations (paper II).
One important issue with regard to our use of previously collected data was that we were not interested in population representativeness per se; we were concerned with specific issues related to valuation methods, regardless of population. Had our focus been on assessing the level of any specific relationship within a certain population, population representativeness would have been crucial.

5.1.2 VALUATION EXPERIMENTS (PAPERS VI AND VII)

5.1.2.1 MODE OF ADMINISTRATION

A market research company administered our surveys. From a standing panel of willing participants, this company drew samples of participants modeled to representative of the Norwegian population on key variables.

We chose to administer online, self-completion web surveys mainly because they cost substantially less than face-to-face surveys. Additionally, the EuroQol group is moving toward a uniform web-based valuation procedure, meaning that future valuation studies are likely to be web-based\(^ {145} \). We also considered the fact that web surveys decrease the burden on participants, in that they do not require participants to travel and also allow participants to complete tasks at any time of the day. The Norwegian population is considered ‘computer-literate’, for instance 92 percent of Norwegian households have internet access at home\(^ {146} \). Norway is therefore a natural candidate for web-surveys.

However, systematic biases in sampling and response rates could be associated with different modes of administration\(^ {147} \). It is probable that our sample is not completely representative of the Norwegian population. For instance, only .5 percent of the respondents in our web surveys were in the 80+ age group, while the corresponding number for the general population is over 5 percent.
We wanted to generalize our findings beyond this specific mode of administration i.e., web surveys, and therefore were concerned with whether different modes of administration could yield different levels of task comprehension or different response strategies. Literature on how modes of administration influence health state valuations is sparse. However, a TTO study comparing an electronic self-completion mode with face-to-face interviews found that the responses to each method did not differ significantly\textsuperscript{148}. Moreover, a computerized web application of the person trade-off elicitation method produced results that were of a similar quality to that produced by face-to-face interviews\textsuperscript{149}.

5.1.2.2 Exclusion Criteria

The exclusion criteria used in preference elicitation studies have been debated\textsuperscript{46,150}. To increase data quality and reduce noise, we chose to use exclusion criteria that were similar to, but slightly stricter than, exclusion criteria that have been applied in most previous TTO-based valuation studies. The assumption that strict exclusion criteria could be beneficial lies in satisficing theory (‘satisfice’ being a combination of the words ‘satisfy’ and ‘suffice’), which is an approach to explaining and characterizing response errors in surveys. Satisficing describes a tendency for respondents to attempt to meet the requirements for adequacy rather than to seek optimal solutions. This theory assumes that the process of integrating information and forming a preference or a judgment involves a considerable cognitive burden\textsuperscript{151}. Some respondents may therefore exhibit satisficing behavior, i.e., may use heuristics to reduce the cognitive burden of a task. Factors that influence the probability of respondents’ satisficing include motivation, ability, and task difficulty. Satisficing behavior manifests itself as higher proportions of ‘missing’ and ‘don’t know’ responses, as well as responses that fail to differentiate between stimuli. Satisficing has been linked to mode of survey administration, indicating that respondents display more satisficing in settings that do not allow non-verbal communication, for
instance telephone or web surveys\textsuperscript{152,153}. Additional satisficing is to be expected with web surveys, because respondents who are at home may be multitasking, reading e-mails, etc. and because web surveys require some level of computer literacy and ability to understand the visual cues rather than individually adjusted explanations from an interacting interviewer\textsuperscript{152}.

We considered the LT-TTO tasks and TTO tasks in our survey to be very cognitively challenging; TTO is usually administered in face-to-face interviews, partly because of its complexity. We therefore expected relatively high degrees of satisficing behavior. To the best of our knowledge, only one published study has compared the pattern of TTO responses elicited on a web application to the pattern of TTO values elicited face-to-face. The primary result of this study was a higher number of values of exactly 1, 0 and -1 (top, middle, and bottom of the TTO scale), in the web TTO values compared to the face-to-face TTO values. Satisficing behavior may partly explain this finding.

Our strict exclusion criteria led to the exclusion of many respondents. However, we consider that the gain in increased data quality and reduced noise justifies the criteria selection.

5.1.3 REGRESSION MODELING

Several of the findings reported in our studies were based on ordinary least-squares (OLS) regression analyses. However, questions pertaining to the appropriateness of using OLS could be raised. For instance, TTO values usually display a characteristic multi-modal distribution with peaks at 0 and 1 and nearly no values in the vicinity of 0. With this distribution of values, residuals from regression modeling rarely achieve a normal distribution\textsuperscript{154}. Our primary aim was to identify the presence or absence of a substantial association between health state values and our variables of interest, for instance sequence number, ATE, or anchoring. Second, we were interested in generating ballpark estimates of the
effects for the purpose of enabling debates about their conceptual importance. To that end, we considered the transparency and accessibility of OLS beneficial; there is evidence that linear models (even when improperly modeled) typically perform well given that the predictors are well selected\textsuperscript{155}. We performed sensitivity analyses to ensure that using more complex analyses did not alter the results substantially. Generally, using more complex regression specifications (e.g., robust regression or including random or fixed effects at the level of individual respondents) resulted in minuscule changes to regression coefficients and standard error estimates. In most cases we did not include these results in our papers, since engaging in detailed descriptions of more elaborate methods could distract readers from our primary message. If the aim had been to control for unwanted effects, more detailed analysis of the relationships would have been required.

TTO values are usually heteroscedastic in relation to severity (i.e., the variance in estimates increases with the severity of the health states), and linear regression does not control for this. Our standard deviation estimates may therefore not accurately represent the precision of our estimates for all levels of severity. However, our conclusions are not crucially dependent on the standard deviations measured within any suitable margin of uncertainty.

Similarly, we chose to be restrictive when it came to including predictor variables and therefore only included variables for which we had a theoretical or empirical reason to expect a statistical relationship.

In most of our papers, we have included graphical presentations of the distributions of modeled values, hopefully enabling readers to make their own judgments regarding the appropriateness (or lack thereof) of our linear regressions.
5.1.4 DISCUSSION OF METHODS BY PAPER

5.1.4.1 PAPER I: COLLAPSED TTO

Paper I’s presentation deviates from our motivation for conducting the study. We started out with several hypotheses regarding MVH TTO that led us to expect substantial noise within BTD and WTD, which made us suspect that the initial choice task would have a greater influence on TTO values than intended. When we devised the cTTO analysis, the idea was to use it as an illustration of problems with the TTO.

The procedure used in the paper is unusual: the resolution of the TTO scale was reduced from a total of 80 possible fTTO values to just 4 cTTO values, and the means of the collapsed scale was compared with means of the full scale. Due to construct overlap, the two sets of mean values could be expected to display relatively high levels of correlation. If the fTTO values were normally distributed around the mean for each health state, the correlation between fTTO and cTTO could be expected to be fairly close to 1. However, the values elicited through the MVH TTO display a characteristic multimodal distribution: for most health states, there are peaks at 0 and 1, smaller peaks at ±.5, and very few values between ±.5 and 0. Exactly what level of correlation should be expected between the health state means of fTTO and cTTO values, given this distribution, is unclear. What is clear is that it should be below 1 by some discernible margin to justify continuing TTO beyond the points captured by cTTO. Separate analyses of mean changes within BTD and within WTD (i.e., of the information that is discarded in the course of fTTO->cTTO transformation) suggested that the observed correlation was caused by a lack of sensitivity to health state severity in WTD and by a very high correlation between mean changes within BTD and mean changes over the full TTO scale.

Is the selected procedure (comparison of cTTO to fTTO) ideal for our purposes? Probably not. As stated in the Objectives chapter, the hypotheses that led us to
our analyses were related to several different effects of how the TTO procedure is performed. A better way to test our hypotheses would have been to tailor specific manipulations of TTO in valuation experiments. However, that approach would have been costly and time consuming, not to mention that it would have been at odds with the objectives of our research project at the time. We therefore tried to devise hypotheses that would be testable with existing data, and the cTTO procedure was the best candidate for analysis.

The reported findings were robust across the Patrick and Torrance transformations, but generalizability beyond those is uncertain.

The findings presented in the paper appear to be robust for MVH TTO-based EQ-5D valuation data across different target populations, respondent weighting schemes, regression model specifications, and transformation methods for WTD states. The uncertainty lies in the interpretation of the findings, and whether the methods used are appropriate for our hypotheses.

5.1.4.2 PAPER II: WTD TRANSFORMATION

The paper compared the magnitudes of impact from switching between UK and US datasets, regression models, and WTD transformation method. Unlike the authors of the US study, we did not impute missing values when estimating our predicted value sets based on the US data. To assess whether our predicted value sets were similar to the published value sets, we compared the published US value set to our corresponding predicted value set. The largest difference for a single health state was .004, and the mean absolute difference for values was .001. Hence, we consider that the predicted value set to be a suitable approximation of the published value set. The published UK value set was identical to our corresponding predicted values.

Generalizability of the relative magnitudes of importance of the three sources of variation is uncertain, because it is highly dependent on the differences in health
state values between the two data sets. However, we were primarily interested in demonstrating that choice of transformation method was important. From our study, we can infer that transformation method is important when estimating EQ-5D tariffs. The impact is a function of the proportion of WTD valuations relative to BTD valuations.

There is no standardized method of directly comparing different value sets. Norman and colleagues\textsuperscript{138} compared different tariffs in several ways, with primary analyses by means of visual comparison of tariff values from different countries similar to figure 8 (section 4.1.2). They made note of absolute levels of values as compared to the UK tariff, and to differences in the magnitudes of the regression constants and interaction terms. Johnson and colleagues performed a direct comparison of US and UK valuation study data, with focus on the mean values for directly valued health states\textsuperscript{137}. They used the Patrick transformation on both datasets, reflecting the fact that using different transformation methods would render the datasets difficult to compare. They then compared each of the measured health states using 2 sample \(t\) tests, and compared the full sets of values by means weighted linear regression. They reported a modeled difference between US and UK values of .095, which is similar to the mean difference in our study of .0925.

We could have used correlations between the estimated values as measures of difference; since the impact of transformation method is a function of the proportion of WTD values, this impact can be described as a linear function of the severity of health states, meaning that if we had used Pearson’s correlation to compare the tariffs transformation method would have had a small impact compared to choice of dataset or regression model. However, a high level of correlation does not mean that using different transformation methods is not important to absolute health state values and gain in estimated HRQoL. We therefore chose to directly compare the estimated tariff values for all health
states and to use the mean of all possible transitions between health states described in the EQ-5D system, because this was similar to the way that value sets are used in CUA. We only used the regression models from the two original studies. Theoretically, model specifications other than the ones we used could have displayed better fit, particularly for the UK data with US transformation and the US data with UK transformation. Because our intention was to rule out the notion that differences between the US and UK regression models were significant contributors to the observed differences between the US and UK tariffs, we did not try to fit any other regression model specifications to our data.

5.1.4.3 PAPER III: LEARNING EFFECTS IN TTO

The only reference we have found in EQ-5D literature on learning effects is an analysis in performed by Golicki et al. on data from the Polish valuation study, in which they compared values from the middle of the experiment (position 6-17) with values from the end of the experiment (position 18-23) in terms of mean and variance. They found no evidence of differences. However, their methods of analysis differed from ours in several ways, particularly in that we performed separate analyses on BTD and WTD values.

The generalizability of learning effects observed beyond EQ-5D valuation using MVH-style TTO is uncertain. Within TTO-based EQ-5D valuation studies, the effects require verification of the findings with other data sets. We recently acquired an MVH dataset that allowed us to determine the sequence number, thereby allowing us to perform similar analyses on UK data. The observed effects were virtually identical to the ones reported in the paper. Health state-specific transfer (order) effects or learning effect interactions with specific dimensions of the EQ-5D description system may exist, but were outside the scope of the study. Our methods of analysis were not suitable for determining the number of
TTO valuations required to maximize response validity; i.e., determining whether later valuations are most valid, as suggested by the discovered preferences hypothesis, or the initial valuations are most valid, as they may be less affected by unwanted framing effects.

Our analyses do not say anything specific about how valuation of specific health states were impacted by learning effects, other than revealing that severe health states with large proportions of WTD values were more subject to learning effects than milder states. We initially performed separate analyses on each of the 42 valued health states. Due to fluctuation in the number of respondents valuing specific health states at specific sequence numbers, combined with individual variations in the values given to all health states, the noise-to-signal ratio made the results at this level of aggregation inconclusive. To reduce the level of noise, we performed analyses of aggregates of all states and groups of health states. Given a much larger sample size, analysis at the level of individual health states could prove highly interesting. A larger sample size could also allow investigation of whether the proportion of respondents valuing specific health states as SWD is a function of sequence number.

5.1.4.4 Paper IV: Ranking vs. TTO

The objective of paper IV was to assess whether ranking and TTO might make respondents sensitive to impairments in different dimensions of the EQ-5D. To this end, we used data from the US and UK valuation studies.

Lacking a gold standard for comparison, several researchers have suggested the use of rank values for comparison. Bleichrodt and Johannson reported that undiscounted MVH TTO values displayed higher correlation with rank order than discounted TTO, SG, or VAS values. However, the analysis was not on the level of EQ-5D dimensions. Most comparisons of valuation methods have not considered potential differences in respondents’ tendency to become more sensitive to
specific dimensions of health, but focus on differences in absolute values, the functional relationship between values from the different instruments, differences in theoretical approach, and differences in propensity to inconsistent responses.\textsuperscript{16,89,160–162}

Due to the differences between TTO and ranking, data obtained using the two methods are not directly comparable; unlike TTO-data, individual rank data does not contain information about the relative distances between ranked health states. One way to circumvent this problem would be to use VAS instead of rank data. However, the MVH protocol uses ranking as a precursor to VAS, such that VAS is performed on health states in ranked order, a procedure that has been found to influence VAS values.\textsuperscript{163,164} We therefore found comparison between TTO and rank data conceptually more appropriate.

Another approach would be to create individual TTO ranks, i.e., order each respondent’s TTO values, and assign the best state a rank of 1, the second best a rank of 2, etc. Using this approach, we could have performed regression analyses on individual rank differences, instead of overall mean rank differences. However, respondents were allowed to assign more than one health state the same rank and were also allowed to give several states identical TTO values. If the tendency to give identical ranks and the tendency to give identical TTO values were different, it would be difficult to devise an unbiased method of calculating rank differences. We therefore chose to reduce the data to mean TTO and mean rank values, and to perform analyses on their respective ranks.

We believe that the strengths of this method lie in its simplicity and accessibility. The primary cost of choosing this method is that the results are very difficult to interpret in terms of absolute size: the regression coefficients indicate the linear difference in mean ranking between the two methods. To get a frame of reference for the observed differences between TTO and ranking, we compared
the rank order correlation between the two methods with correlations between
the US and UK data using the same method.

Within the universe of MVH-protocol-based EQ-5D valuation studies, we think
that our methods are suitable for identifying the presence, but not the
magnitude, of dimension-specific differences in sensitivity between rank and TTO.

5.1.4.5 Paper V: Experienced and Hypothetical Health

The objective of the study was to identify and illustrate differences in the
patterns of relative importance attributed to the dimensions of the EQ-5D by the
general population (hypothetical health, HH), and by respondents valuing their
own health (experienced health, EH).

Data from MEPS were selected for the EH part of this study, because MEPS was
the largest set of data we could find with self-ratings on the EQ-5D that were
combined with a more holistic measure of experienced HRQoL, in this case the
VAS. Other large EQ-5D patient datasets are available, but several (for instance
the Health Survey for England) do not have matched EQ-5D and VAS/other
HRQoL-data, a requirement for our purposes.

In the MEPS setting, VAS ratings were done for one state only: the state
experienced by the respondent. These differences in setting make direct
comparison difficult and are part of the reason why we argue that comparisons
between EH and HH should not be performed in terms of absolute values.

To render EH and HH values more comparable, we devised a set of procedures
that allowed us to focus solely on differences in dimension-specific weights within
each of the respondent groups. We then chose to present the patterns of relative
importance observed within the two groups without performing formal statistical
tests comparing them.
All of the steps used to render the presentation of relative dimension importance could have been performed in numerous other ways. Sensitivity studies varying aspects like the order of the standardization and adjustments performed, and altering the regression model indicated that the results were quite robust to such changes.

By intention, the methods chosen deviated from the ones used by Mann and colleagues\textsuperscript{28}: they compared each dummy variable coefficient, for example $u_3$ (representing extreme problems on the usual activities dimension), between EH and HH in terms of their absolute magnitudes. As can be seen from figure 10 in section 4.1.5, $u_3$ was the level 3 dummy displaying smallest difference between our EH and HH models in terms of absolute values. However, in terms of the relative importance within group, the usual activities dimension was the most important in the EH group, and the least in the HH group.

Similar to our method of comparison, Norman and colleagues report differences between the Japanese and the UK tariffs in terms of the relative magnitudes of different dimensions, pointing out that the Japanese population apparently adhered relatively greater importance to the mobility dimension, and less importance to pain/discomfort and anxiety/depression\textsuperscript{138}.

5.1.4.6 PAPER VI: ATTITUDES TOWARD EUTHANASIA

Attitudes toward euthanasia could be part of a larger construct entailing attitudes toward or acceptance of death in general, not only in a euthanasia setting. Euthanasia attitudes could be composed of several sub-constructs that more or less overlap with ATE –adherence to religion, fear of death, etc. In addition, we used a non-standardized, rather crude measure for respondents’ ATE.

In our analyses, we calculated ATE scores as the mean of three 5-point Likert-style items, where the categories labeled “strongly disagree”, “disagree”, “neither disagree nor agree”, “agree”, and “strongly agree” were assigned equidistant
values ranging from -2 to 2. The three items were statements phrased to measure attitudes toward active euthanasia, passive euthanasia, and assisted suicide. Handling these as linear is based on two assumptions, namely that the three items (at least to some extent) measure the same underlying construct, and that the relative distance in “severity” between the different steps are similar. The assumption of common construct was supported by the framing of the task: the three items referred to the same presented case, were phrased similarly, and were strongly related, conceptually. The Chronbach’s alpha observed (.76) was considered satisfactory, since alpha is a positive function of the number of items. The assumption of linearity with regard to the aggregated Likert-type scale is based on generalized linear item response theory, and linearity is often assumed of Likert-type scales with the specific set of labels, as seen for instance in the personality questionnaire NEO-PI-R. Granted, there is considerable disagreement in the literature regarding how Likert-type scale data should be used. However, while such scales may not adhere perfectly to the requirements of such analyses, there is evidence supporting the use of linear methods in cases where the breach of requirements are not too profound. While each item may at best be ordinal in nature, the aggregate of several items may be assumed to have interval properties. An alternative approach using 14 dummy variables to assess the specific movements in health state values for all steps along the ATE scale would reduce statistical power to the point where we would require substantially larger respondent samples than we had available.

All respondents answered the ATE questions after completing the valuation exercises, which means that they could have been influenced by the previous LT-TTO or TTO tasks when they answered the ATE questions. It is also problematic that respondents may have interpreted the TTO question as a covert attempt to elicit preferences for euthanasia, which would imply that we measured the same construct twice. Our findings should be confirmed with other data sets to assess their generalizability.
5.1.4.7 Paper VII: Anchoring bias in TTO

To the best of our knowledge, anchoring from the starting point of the task has not been observed in TTO settings previously for the simple reason that the starting point has typically been fixed. However, there is evidence that TTO values are sensitive to the length of time presented (which has typically been fixed at 10 years), and that LT-TTO is sensitive to the length of the lead-time\textsuperscript{171}.

As with several of the other papers, the non-normal distribution of TTO values complicated decisions regarding what type of regression model would be most appropriate for this study. As a form of sensitivity analysis, we tested different options, including OLS, versions with random effects at the level of individual respondents, and robust variance estimation. All tested variants identified the TTO starting point value as the only statistically significant predictor. Since the robust regression relaxes some of the requirements of OLS that are not fulfilled by TTO values, we presented results from this model in the paper. Selecting other variants would not alter any conclusions.

Primary analyses were performed on LT-TTO with 11 different starting points, while MVH-style TTO was compared to TTO5. The magnitude of mean difference between MVH TTO and TTO5 after exclusions (.085) was smaller than the magnitude of a similar starting point movement (.5) in the LT-TTO. Before exclusions, the observed difference was much greater, but may be due to a large proportion of respondents in the MVH group exhibiting satisficing behavior, as indicated by the tendency to trigger exclusion criteria. Generalizability from the LT-TTO study to MVH TTO is complicated by the ping-pong routing of the MVH protocol, since altering the starting point inevitably alters the routing procedure. The gap effect observed in MVH TTO was not found in the TTO5 group, suggesting that it is specific to the starting point of the MVH protocol, but also making the two groups of values less comparable.
The rate of respondents triggering the exclusion criteria was strongly related to the starting point. As one of the criteria was related to WTD values (all values <=0), this was not completely unexpected. However, if the number of WTD values were to trigger the exclusion criterion because of an anchoring effect, excluding these respondents would lead to underestimation of the anchoring effect. On the other hand, the rate of respondents triggering the other two criteria was also related to lower starting points, possibly indicating that low starting point increased the difficulty of the task. Since the exclusion criteria were biased in terms of the target variable (starting point value), we opted to report our findings both with and without exclusion to allow readers to form their own opinion.

5.2 GENERAL DISCUSSION

The structure of this section diverges from the previous ones in that it is not organized primarily by paper. We start with a discussion of influence from research choices and construct-irrelevant factors on TTO valuation of hypothetical health states. Results from five papers (papers I, II, III, VI, and VII) are discussed in this start section. From there we move on to more general considerations regarding implications of these and other findings: uncertainty of TTO values in terms of their absolute relation to the anchors “perfect health” and “death”, implications for the initial choice task of MVH TTO, handling of WTD values, and a discussion of the challenges related to valuing death.

After these topics, related primarily to TTO values, we consider the inconsistency observed between ranking and TTO (paper IV). We then present a critical perspective on the legitimacy of EQ-5D tariffs interpreted as population-based. Finally, we consider the differences observed between how general population respondents and patients consider the relative importance of the different EQ-5D dimensions (paper V).
5.2.1 INFLUENCE OF RESEARCH CHOICES AND CONSTRUCT-IRRELEVANT FACTORS

Our studies demonstrate how valuations of EQ-5D states are influenced by several factors of questionable relevance and show that the choices researchers make when designing valuation studies impact on estimated EQ-5D tariffs.

Whether an effect is problematic depends on the construct one wishes to measure. In estimating HRQoL for use in QALY analyses, some factors are clearly irrelevant, while the relevance of other factors may be subject to diverging opinions. When estimating the importance of a problem, the magnitude of its influence must also be considered: A large effect of something that is considered clearly construct-irrelevant is very problematic while a small effect of a factor that might be construct-irrelevant is less of troubling. Thus, two judgments are required: one regarding the degree of construct-relevance, and one regarding the magnitude of influence. Considerations of construct-relevance entail debating theoretical requirements and the intended interpretation of the measure. Magnitude judgments require some metric of comparison to determine the threshold at which a magnitude goes from being of little concern to being important or substantial. However, the influences we consider are of different types, and we know of no standardized method of magnitude comparison in the context of EQ-5D tariff generation.

The closest thing we have found to a threshold value for importance is the concept of Minimal Clinically Important Differences (MCID, also referred to as Minimally Important Difference, or MID), which may be defined as “The smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient’s management”\textsuperscript{172}. As with the measurement of tariff values, there are numerous methodological challenges in determining MCID estimates, but we consider these outside the scope of this
thesis. A study by Walters and Brazier estimated the mean MCID using the UK EQ-5D tariff to .074\textsuperscript{173}. Pickard, Neary, and Cella found MCID values in cancer patients of .08 using the UK tariff and .06 using the US tariff\textsuperscript{174}. The MCID may be tariff dependent, since different tariffs span different value ranges. With UK EQ-5D tariff values ranging from 1 to -.59 and US values ranging from 1 to -.11, the mean MCID estimate suggests that a change of roughly 5% of the tariff range may be considered important. Where possible, we will use 5% of the tariff range as a ballpark threshold for considering an influence magnitude important.

Construct-irrelevant factors are usually defined as irrelevant factors that influence the measurement of a target construct. Several of our studies examine how factors may influence the measurement of health state values when using the TTO. According to the Standards for Educational and Psychological Testing\textsuperscript{91}, the validity of any measurement is directly related to the intended interpretation of the measure. In the case of EQ-5D valuation, the intended interpretation is related to the intended use of the resulting health state values, i.e., to their use as tariff values for the purpose of estimating gains and losses of HRQoL. From this perspective, all steps influencing final tariff values should be evaluated in terms of their contribution to or distortion of resulting tariff values and how they may influence subsequent priority settings. In addition to the valuation procedure itself, final tariff estimates are also influenced by other methodological choices on part of the researchers, including procedures such as (but not limited to) the selection of reference group, WTD transformation, and regression modeling. We will therefore adopt the perspective that all steps involved in tariff generation may be subjected to analysis of their construct relevance.

5.2.1.1 WTD TRANSFORMATION —PAPER II

The choice of transformation method for WTD states is a post-hoc modification of values. The motivation for transformation seems to be to achieve mean health state values that are within an acceptable range, since a tariff based on
untransformed values would lack face validity and risk not being used. Since transformation methods lack theoretical and empirical support\textsuperscript{69,123}, we consider the choice of transformation method to be essentially arbitrary and completely unrelated to HRQoL, thus highly construct-irrelevant. We have demonstrated that the choice of transformation method may have greater impact on the tariffs than switching target population. The exact magnitude of effect is not directly comparable to our chosen 5% threshold of importance; the effect varies greatly by health state, and it is not obvious exactly how the magnitude should be calculated. However, the observed mean difference between estimated tariffs using the two different transformations (absolute value of 0.137) is greater than reported measures of MCID for EQ-5D\textsuperscript{173,174}. If the post-transformation tariffs were to be compared to mean-based tariffs without prior transformation, the difference would be of much greater magnitude. Therefore, we consider the magnitude of effect to be substantial. Since transformation methods are construct-irrelevant (in that they do not measure severity of health states), and considering the importance of the choice, we conclude that they pose a substantial problem.

The choice of transformation method (if we limit our selection to the Patrick and Torrance transformations) has substantial impact on the slope or steepness of the tariff. This has consequences for how many health states end up with estimated WTD values, and for the magnitude of QALY-differences incurred by health interventions\textsuperscript{138,175}. Unlike the tariff values, the ordering and relative distance between health states is not greatly influenced by switching the method of transformation, suggesting that this choice has limited impact on the interpretability of TTO-based tariffs in terms of relative severity of different impaired health states.
5.2.1.2 ANCHORING – PAPER VII

The starting point in the valuation task is unrelated to the severity of the target health state. A central assumption underlying choice-based methods like TTO is that such methods are suitable for identifying, but not influencing, the point of preferential indifference between options offered. Thus, the starting point of the task should have no influence on the resulting values. It follows that any observed influence on the part of the starting point would constitute a breach of procedural invariance. Clearly, tariff values for health states should not be determined by the starting point in TTO. We therefore consider the starting point construct-irrelevant to the elicitation of HRQoL values.

Observed anchoring directly affects the absolute values of health states, rendering their interpretation in relation to the QALY scale uncertain. The magnitude of the observed effect is large (a movement of the starting point value of 1.0 resulted in a mean change in elicited values of .25). A mean effect of .25 on all health states on a scale theoretically ranging from 1 to -1 is greater than our chosen 5% threshold of importance.

Since there is no theoretically correct starting point, anchoring may be unavoidable. We consider this a substantial problem for the validity of TTO values in terms of their interpretability in relation to the QALY-anchors of ‘perfect health’ and ‘death’. However, if anchoring is linear, it may still be possible to interpret TTO values in terms of the relative distance between different impaired health states.

5.2.1.3 LEARNING EFFECTS – PAPER III

Principally, we find worrisome the implication that the number of valued states in a valuation study should impact on the resulting tariff values, as the number of health states valued in a study is irrelevant to HRQoL. However, if the learning effect is related to improved understanding of the task, as comes with increasing
experience, earlier valuations should be considered biased due to construct-underrepresentation. Conversely, if the learning effect is related to increased influence of framing effects in the task, for instance, the effects of invoking ‘death’, later values may be increasingly influenced by construct-irrelevant factors. Compared to the findings in our other papers, we consider the magnitude of the effect observed to be relatively small. As long as the learning effects are unrelated to specific dimensions of health, the problem is limited to increased uncertainty with regard to the absolute values for all health states. The observed magnitude of our findings would suggest limited net effect on resulting tariffs. However, the fact that respondents seem to grow increasingly averse to eliciting values in the vicinity of death (0), interpreted as exacerbation of the gap effect may suggest that the observed effects stem from framing introduced by direct comparison to death. If the learning effects are a consequence of repeatedly reinforced framing effects, an important question is how much base-line influence may be present from the very first valuation.

Attema and colleagues present a different interpretation of the gap effect, suggesting that it is an artifact created by respondents’ tendency to elicit values that require fewer choice task iterations - a form of satisficing behavior. They argue against the incremental routing procedure used in the MVH TTO protocol after the initial jump from comparison to death to comparison with 5 years in perfect health, since this procedure requires relatively many iterations to reach the TTO values -.1 and .1. Similarly, the observed learning effects could stem from increased satisficing. However, there are two reasons why we do not believe this interpretation to be correct: the MVH TTO protocol requires as many steps to reach values of -.9 and .9 as it requires to reach -.1 and .1. Yet there is little evidence of tendencies not to elicit these values (see figures 7 in section 4.1.1, and figure 13 in section 4.1.7). And we found no indication of respondents eliciting values requiring fewer choice tasks with increased valuation experience.
5.2.1.4 Attitudes Toward Euthanasia – Paper VI

ATE appears to influence TTO and LT-TTO values, but not VAS values, regardless of how VAS values are rescaled. Whether or not attitudes toward death, and more specifically, attitudes toward euthanasia, should be construct-irrelevant is a complicated issue. The debate relates to fundamental issues regarding what factors the QALY weights should include, whether impaired health and ‘death’ are commensurable units, and whether ‘death’ can be valued at all. A more extensive discussion of this topic follows in section 5.2.5. The magnitude of the effect is substantial: an increase of 1 on our ATE scale (which ranges from -2 to 2) is associated with a mean change in MVH TTO values of -.113, greater than reported MCID measures for EQ-5D. The absolute level of elicited health state values seems to depend on the salience of ‘death’ in the valuation task, which emphasizes the importance of the debate about which aspects about death are relevant when valuing impaired health states.

5.2.1.5 Collapsed TTO – Paper I

The findings of paper I analyzed in conjunction with findings of our other papers, indicate that respondents were influenced by characteristics of the initial choice task (see initial choice task). These may have included communicative implicatures of presenting three categories and influence of invoking ‘death’. To the best of our understanding, the choice tasks in TTO were not in any way intended as instructions regarding what values respondents should assign to health states. Therefore, any impact on elicited values by respondents’ interpretations of unintended implicatures in the framing of the task should be considered construct-irrelevant. For more elaborate discussion of whether influence of invoking death in the valuation task should be considered construct-irrelevant or, see section 5.2.5.

We believe that cTTO, as described in paper I cannot be interpreted as a valid stand-alone measure of HRQoL. From our analyses, it is difficult to estimate the
magnitude of influence from the initial choice task on the tariff. However, the lack of any new information stemming from continued valuation suggests that the influence of the initial choice is greater than intended, or that respondents did not make trade-offs as intended.

5.2.2 Absolute Values

Within the QALY framework, the absolute, and not just the relative values must be valid. In the words of Jeffrey Richardson

To provide a ‘valid’ utility score — a number which truly represents what it purports to represent (in this case, ‘utility’) — the numbers produced by each of these techniques must satisfy several demanding conditions. First, they must possess an ‘interval’ property in the conventional sense. This implies that an increase in the utility score from 0.2 to 0.4 has the same meaning as a move from 0.7 to 0.9. This was described by Richardson as the ‘weak interval property’. Second, the utility scores must have a ‘strong interval property’ (Richardson 1994). This implies that, for example, a 10% increase in the index of utility from, say, 0.7 to 0.77 has exactly the same impact upon utility as a 10% increase in the life years obtained from a project: for example, an increase from 20 to 22 life years. Third, following from both of these requirements, the absolute and not just the relative value of utility numbers must be a valid representation of utility. This third property cannot be directly observed since you cannot demand that people live out their stated preferences.

WTD transformation has substantial influence on the absolute scale of the resulting tariffs (i.e. the relationship between health state values and the anchors ‘perfect health’ and ‘death’, but not on the relative distance between impaired health states). The observed learning effects imply that the slope of the tariff is a function of the number of health states valued by each respondent. The observed anchoring effects suggest that the starting point of the TTO task directly influences the absolute level of resulting tariff values. Similarly, the relationship between attitudes toward euthanasia and TTO values suggests that varying the salience of ‘death’ in the valuation task influences the absolute level of resulting values. Combined, these findings suggest that while the relative distance between different health states may be fairly stable across construct-irrelevant...
methodological variations, the absolute level of the values varies considerably. The magnitude of the uncertainty associated with the absolute scale of TTO values that is due to these factors is difficult to pinpoint since we have only tested a fraction of all possible TTO variants\textsuperscript{103,171,178}.

It seems possible to manipulate the absolute level of elicited TTO values by altering for instance the starting point. Given the possibility that similar manipulations could be performed with SG, it should be possible to identify variants of TTO and variants of SG that display agreement in terms of absolute levels of their respective health state estimates. However, this would only accomplish a splendid example of the susceptibility of TTO, and possibly SG, to manipulation.

The uncertainty with regard to the absolute scale is at least equal to the observed mean difference between the tested extremes in the anchoring experiment. This has implications for the validity of TTO values as weights in QALY-calculations, since the statistical uncertainty of health state estimates does not account for the much greater underlying uncertainty in the method.

5.2.3 THE INITIAL CHOICE TASK

Several of our hypotheses were related to the initial choice task in MVH TTO. Ensuing analyses (paper I) revealed that knowledge of the crude categorization of health states by each respondent was sufficient to predict resulting tariff values. Furthermore, we found that values elicited with the WTD method were unable to discriminate between good and bad health states. Both these findings were in line with our hypotheses, but they do not inform us of the reasons behind the observed effects. Thus, paper I alone does not yield much information as to which of our hypotheses regarding the initial question could be at work, and to what extent.
Considered together with the other papers, some interesting patterns emerge: The distributions of values elicited using different TTO methods (paper VII) suggest that the gap effect is not observable in LT-TTO. This is in line with findings by Devlin et. al.72. Similarly, starting the MVH task at 5 years of perfect health seems to completely eliminate the gap effect in the BTD range and, to some extent in the WTD range. This suggests that the initial choice task of MVH-style TTO may have an effect that is unique compared to other possible starting points (the question is encountered in TTO 5 if the respondent “spends” all time in perfect health). This, and the finding that respondents distance themselves increasingly from values surrounding ‘death’ (learning effects observed in paper II) may be interpreted as supporting the idea that respondents are discomfited by the comparison to immediate death. The interpretation that this may stem from effects of invoking death is supported by our findings regarding attitudes toward euthanasia (paper VI).

Interpreted as a request for sorting some health states into each of the initial categories (see the section 1.6.2 on communicative implicatures), the starting point of the MVH protocol may inflate the number of states valued as equal to or worse than death; despite the fact that ETD is a point, and while BTD and WTD are ranges of values, respondents may feel compelled to assign some of the health states to each of the initially-presented categories. In accordance with this view, the starting-point graph from paper VII (Figure 13, section 4.1.7) illustrates a tendency for respondents to state preferential indifference for health states at the starting point. In the same figure, a strong relationship between the starting point of the LT-TTO and the number of WTD valuations may also be observed. However, the number of values at precisely 0 actually declined as starting points were lowered, with the exception of the group that started at 0, suggesting that the initial choice task had effects in line with respondents adhering to the principle of cooperation, in addition to anchoring as described by Kahneman and Tversky (see section 1.6.1.2). Consider, if only anchoring was at work, one would
expect the tendency to elicit any specific value to be a positive function of proximity with the starting point.

If the specific starting point chosen in the MVH protocol is construed as a request for respondents to assign health states to ETD and WTD, the elicited values cannot be interpreted in relation to ‘death’ with any certainty. Inflation in the number of states considered WTD would also contribute to the problem that led to the need for post-hoc transformation methods.

Combined, the findings from paper I (suggesting that respondents are influenced by communicative implicatures), paper III (suggesting that invoking death increases distancing from death), paper VI (supporting effects of invoking death), and paper VII (illustrating anchoring effects and supporting influence from communicative implicatures) suggest that all possible starting points for the TTO will influence resulting values. Since the gap effect appears to be specific to the MVH protocol, the starting point used may be particularly problematic.

5.2.4 WORSE THAN DEATH VALUES

The incorporation of WTD values in TTO was done by Torrance in 1982\textsuperscript{179}, following several studies indicating that there were cases in which some respondents would consider death preferable to described health states. He transformed WTD values using what we refer to as the Torrance transformation stating that “This maintains symmetry between the value scale for states preferred to death which runs from 0 to 1, and the scale for states dispreferred to death which runs from 0 to -1.”\textsuperscript{5} However, he later cautioned that “… the cardinal utilities must be treated as preliminary since the validity of the time trade-off measurement technique for states worse than death has yet to be established.”, and that “In summary, the measurement of cardinal preferences for health states worse than death is still at a very early, primitive stage. Much more study is required before reliable valid instruments will be available.”\textsuperscript{64} Until such evidence
is available, he suggested that WTD values be segregated in reports, rather than to “bury them in the calculations.” He suggested performing extensive sensitivity analysis on the preference values, varying the value of WTD states from 0 to some large negative value to illustrate their impact on resulting preference estimates.

Our papers shed light on the validity of WTD TTO values from different angles. The aim when performing a TTO valuation study is to value how good or bad health states are on a cardinal scale. For this purpose, one must have a valuation method that enables respondents to discriminate between good and bad health states. However, paper I revealed that means of WTD values for the 42 directly valued states were stable across severity in both the UK and US valuation studies. This finding could be interpreted in different ways: For instance, we consider the WTD task to be complex and cognitively demanding. If a large proportion of respondents misunderstood the task, the distribution could be dominated by noise. On the other hand, respondents may have understood the task, but were unable to, in a meaningful way, quantify how much worse than ‘death’ a health state was, indicating a problem of commensurability between the value for health states and death (see section 5.2.5). The WTD task entails comparisons with ‘immediate death’ for each iteration. If invoking death has adverse effects on how respondents perform valuation, it is likely that the WTD values will be much more affected by such effects, increasing the noise to signal ratio.

Since WTD values apparently fail to discriminate between health states of different severity, the contribution of WTD values to tariff values is restricted to a semi-fixed effect (there is some noise) that is determined by the proportion of respondents falling into the category and the chosen method of WTD transformation.

While WTD values were unrelated to the severity of the health states, they changed substantially as a function of the number of previously performed TTO
tasks (paper III). Empirical studies have concluded that uncertainty about preferences increases the influence of framing effects. The presence of substantial learning effects on WTD values could therefore be taken to indicate that respondents are very uncertain about these values.

Regardless of the cause, our findings suggest that there is little practical point to continuing the valuation task once respondents have categorized the health state as WTD. Furthermore, the continued usage of theoretically lacking post-hoc transformation and the variety of suggested mathematical methods for handling the problem caused by these values (median modeling, circular regression, lead/lag-time TTO, and other solutions) suggest that investigators have limited faith in the validity of the elicited WTD values.

5.2.5 THE VALUE OF ‘DEATH’

The debate regarding the value of death is old, dating back at least to Aristotle and Epicurus; Aristotle argued that death is the worst thing that can happen to man, while Epicurus argued that since death is the end of conscience, death precludes experience, and therefore death cannot have any intrinsic value; it is neither good nor bad.

As we understand it, the QALY does not require death itself to have a value, since death implies that there is no time spent alive. The QALY is the area under the quality of life curve as a function of time. Given ‘death’ conceptualized as ‘no time spent alive’, the accrued QALYS will always be 0, regardless of the QALY weight assigned to death. Consider the standard graphical representation of the QALY system (figure 14). If x (<= 1) is the QALY weight, a 1 year life with a QALY weight of x is equivalent to a life of x years in perfect health. As x approaches 0, life in perfect health equivalent to 1 year with a QALY weight of x approaches infinitely short length. From this perspective, any numbers of years in a state with a QALY-weight of 0 equals a life of 0 length, i.e., not being alive.
Figure 14 - The QALY equivalence

When TTO was first devised, it was restricted to the BTD scale. Following several studies reporting that some respondents indicated a preference for death to some presented health states\textsuperscript{62,63}, Torrance adjusted the method to allow valuation of WTD states\textsuperscript{179}, but pointed out that “Much more study is required before reliable valid instruments will be available”\textsuperscript{64}. One of the first valuation studies eliciting negative values with the TTO was designed by Patrick et al\textsuperscript{65}. The authors acknowledged that it could be cognitively demanding to perform comparisons to death, and that it could be difficult for respondents to estimate cardinal values for WTD states. They also noted that respondents could be influenced by vastly diverging aspects of death and manner of dying: religious beliefs, previous experience, and other factors. Common for both Patrick and Torrance is that they do not question whether WTD valuation is meaningful, or
reflect upon which aspects of death should be considered construct-relevant in valuation of health.

TTO may seem like a direct implementation of the QALY. However, unlike the QALY, the TTO task (at least for values <=0) implies that there is a concrete value for death. Using ‘death’ as an anchor, the procedure involves comparing life in impaired health to death. To make these comparisons, ‘death’ requires a value. The validity of asking respondents to value ‘death’ is rarely questioned. Since death is a biological certainty, some argue that death is an appropriate comparator for health state valuation, as exemplified by Patrick Sullivan in a recent editorial in Medical Decision Making:

> Within the context of QALYs, selecting reference points at death and full health is particularly convenient and meaningful. Assigning a value of 0.0 for death ensures that when a patient enters the death state, QALYs cease to accrue. [...] Although the anchor of death (0.0) is measurable and concrete, the description of “full health” is much more nebulous and will vary across descriptive systems.185

However, concluding, on the basis of death as a biological certainty, that the psychological value attributed to death is equally certain is a logical fallacy. Generally, there is no mechanism through which a biological certainty of measurement related to a state (e.g., determining if someone is dead) translates into certainty regarding psychological valuation of said state. Furthermore, neither the ability to ask a question, nor the fact that people respond to it, constitutes sufficient evidence that the question is meaningful. In accordance with the cooperative principle (see section 1.6.2 on communicative implicatures), asking a question implies that there is meaning involved. Lacking apparent meaning, respondents will do their best to interpret the question in a way that makes the act of answering meaningful, as they will try to fulfill the requirements of the interviewer. This means that having people choose between a set of available options does not necessarily imply, in any factual sense, that the options are commensurable, or even comparable. Thus, comparing living states involving
various degrees of health impairment to ‘death’ is not necessarily meaningful, and does not necessarily allow the extrapolation that health states can have meaningful values in any specific relation to death. In the words of Ezra Hauer: “People may be willing to express preferences and econometricians may be eager to interpret them. But, inasmuch as these preferences are vacuous, they have no interpretation”186.

Following the Epicurian view, death is not a state, but rather the absence of all possible states of being. Exempt from religious beliefs about what happens after death, valuation of death thus consists entirely of attributes associated with death, but not death (or “being dead”) per se; pain to loved ones, fear of the unknown, the process of dying, missed opportunities etc. To make sense of comparing impaired health to death, respondents have to value these death-related attributes. Unfortunately, since respondents are not instructed as to what aspects related to death they should consider when performing valuation, different respondents are likely to emphasize different aspects in valuation of death. The net contribution of each of these death-related attributes is unknown. However, if respondents do anchor their responses in different constructs, the validity of aggregation of such values may be questionable.

We have not been able to determine from existing literature exactly which death-associated attributes, if any, should be considered construct-relevant in TTO valuation. Macran and Kind argue that one should avoid wordings that lead the respondents to focus on the process of dying123, indicating that the authors considered this process construct-irrelevant. This may be in contrast with Patrick et al65, who mention the process of dying as one of several aspects that respondents may consider, without making any statements regarding the relevance of each aspect. According to Fowler, a respondent characteristic ‘reluctance to give up life’ is confounded with the value of the health state when valuation methods involve death or risk of dying176. The implication is that the
authors consider ‘reluctance to give up life’ to be construct-irrelevant when measuring health state values. In an alternative TTO procedure, Buckingham, Birdsall, and Douglas operationalized trading time as the need for more sleep (the number of hours an individual would accept of additional sleep per day to avoid an impaired health state)\(^{187}\). The authors advocate that the advantage of this method is to avoid triggering fear of death, leading us to interpret that they considered fear of death as construct-irrelevant. They also mention that including death in the valuation study is problematic, for instance because some respondents would include religious beliefs about reuniting with deceased loved ones. This suggests that they consider beliefs regarding what happens after death to be construct-irrelevant.

In accordance with the definition of validity used in this thesis\(^{91}\), determining what is construct-relevant in the TTO task should be guided by the intended interpretation of the measure, i.e., the use of TTO values as QALY weights for health states. As mentioned, respondents may be influenced by several death-associated attributes when confronted with death in TTO valuation. We will consider some of these in more detail.

Considerations of the process of dying may influence some respondents. However, we would argue that this is not an aspect that QALY weights is intended to measure, since the process of dying would necessarily entail deviation from the presented health state. Such deviation from the concept of measurement implies construct-irrelevance.

A similar argument may be used for fear of death; this is something that happens while alive. If valuation of ‘death’ is influenced substantially by considerations like fear of death, death itself is not valued, but a negative aspect of life is.

Pain to family and loved ones is not an aspect of any health state, but a possible external consequence. To the best of our understanding, the QALY system is
intended to measure qualities related to individual health, not impact on others. If this applies to health states, it should apply to death, or at least a sound argument should be presented justifying why death is to be treated differently.

Valuing ‘death’ as the loss of opportunity is also troubling, since this makes the value assigned to ‘death’ a function of the respondents’ perceived likely future life. Arguably, the same problem exists for comparisons with finite lives in any health state, since respondents may conceptualize the value of $x$ years in perfect health to be the loss of opportunity for the perceived likely future life following the $x$ years. However, asking respondents to consider a life of $x$ years in perfect health is more likely to induce a framing where respondents focus on the positive value of these years than is having them consider the concept of ‘immediate death’ (see section 1.6.1.1 on framing effects). Again, if the QALY ‘death’ is conceptualized as a life of 0 length (not being alive), losses related to the time after death fall outside the scope of interest.

A helpful perspective may be considering the relationship between the QALY and time before birth. Life has two end points, birth (or conception, according to some), and death. In the finite period of time between these two end points, people are in health states, and they accrue QALYs. There is no conceptual distinction in the QALY system between time before birth and time after death, since both entail ‘not being alive’. Unless one introduces a conceptual difference between time after death and time before birth, any concept irrelevant to the value of before-birth-time should also be considered irrelevant to death. To us, this implies that whatever happens after death is irrelevant to the valuation of health states.

From this perspective, religious considerations, including the belief in life after death, heaven, hell, reuniting with loved ones etc., should be considered construct-irrelevant in health state valuation. If respondents were asked to compare health states to not yet being born or not being born at all, we assume
that they would be less influenced by concepts such as fear of death, religious outlooks, or considerations of the process of dying. Unfortunately, it is uncertain if, after removing all irrelevant factors (i.e., considering the value of ‘not being born’), there would be anything left with which it would be possible to compare the value of health states.

Our reservations regarding the use of methods that require valuation of ‘death’ are far from unique or new: in other fields of economic valuation involving health risks, for instance, traffic safety, the common method for valuing life is a specific kind of willingness to pay (WTP) where respondents are asked to state their WTP for small reductions in risk of dying. The use of this procedure is usually attributed to Schelling, who argued specifically for the impossibility of valuing ‘death’ directly. Schelling claimed that invocation of ‘death’ triggers death anxiety, which ruins the chances of rational valuation of death itself. According to Schelling, it is possible to indirectly measure the construct-relevant factors related to ‘death’ using the marginal death-risk WTP procedure. The rationale is that there may exist a sweet spot, in terms of marginal death risk, within which respondents are able to quantify their perception of the disvalue of a death risk, while death anxiety is not triggered to the point where it impairs judgment. Too low probabilities of death would be impossible for respondents to conceptualize, while too high probabilities would lead to useless values due to irrelevant factors, including anxiety. Thus, Schelling, and the substantial body of subsequent literature supporting WTP using marginal reductions in risk of death, is at odds with the basic premises of both TTO and SG. Exactly which factors related to death remain to be valued, given that the sweet spot was found, were not stated explicitly by Schelling. Others have argued that there is insufficient evidence to support the notion that the WTP task measures a meaningful construct, rather than a spurious number that results from the way in which the task is presented. They point out that valuing any specified risk of death requires a preceding
valuation of death\textsuperscript{186}, and that WTP estimates resulting from this procedure display enormous variance\textsuperscript{189}.

To summarize: The debate regarding the value of death is old and ongoing in philosophy, but very limited in the literature on EQ-5D valuation. The QALY does not necessarily require death to have a specific value or to be valued directly, and it is uncertain if death can be valued at all. TTO requires valuing ‘death’, but it is unclear what aspects related to ‘death’ should be included in such valuation. Since respondents are not instructed as to which aspects of ‘death’ they should focus on, it is unknown what they actually do consider. Our findings suggest that attitudes toward euthanasia may be one such factor (paper VI) and that the initial choice task in the MVH TTO has considerable impact on how people value health states (paper I, paper III, and the existence of the gap effect\textsuperscript{133}), and the distinct differences in the distribution of TTO values between MVH TTO and LT-TTO.

In conclusion; we do not agree with Patrick and Torrance that observation of respondents valuing health states as being WTD is sufficient evidence that this is meaningful in the context of QALYs. However, we share the worry of Torrance that such values may not necessarily be measured with validity\textsuperscript{64}. A core issue is the lack of a clear definition of what ‘death’ means in the QALY context. We would argue that ‘death’ should be defined as equal to ‘not being alive’ or ‘not being born’. However, regardless of the chosen definition, valuation tasks should inform respondents of what aspects of death (if any) they should consider.

5.2.6 Rank inconsistencies

Thus far, we have focused on findings related to TTO. However, paper IV shows that respondents performing TTO and rankings of the same health states display a tendency to rate health states with impairments on the three first EQ-5D dimensions (mobility, self-care, and usual activities) as relatively worse when
performing ranking than when performing TTO. The finding that different methods yield different values is far from new, and was pointed out by Patrick:

“An underlying assumption of health-state-preference measurement is that an individual has relative preferences that are consistent, at least within methods. Our results add to the literature that shows different preferences are obtained for the same health state depending on the method of measurement.”

However, most previous comparisons of different valuation tasks have been concerned with the functional relationship between them – for instance concluding that TTO generally results in lower estimates than SG, and that VAS results in lower estimates than TTO. As several of our papers illustrate, the absolute values of health state estimates elicited using TTO seem to be uncertain and to depend to a substantial degree on construct-irrelevant factors (see section 5.2.2).

Disagreement between instruments or variants of instruments with regard to the absolute level of health state measures is less troubling than disagreement between the relative weights of the different dimensions of health. With two specific exceptions, the ranking of health care interventions by cost per QALY are determined by the relative weights of the different health impairments in the tariff. In other words, the ranking of interventions is insensitive to the absolute scale of the tariff values used. The exceptions to this rule are cases involving movements between impaired health states and ‘perfect health’ or death, both of which are sensitive to the absolute scale of tariff values. The findings in paper IV imply that using tariffs based on ranking would result in relatively greater resource expenditure on interventions that improve the dimensions mobility, self-care, and usual activities than would tariffs based on TTO.

At worst, this finding implies that not only is the absolute scale of QALY weights based on TTO uncertain, but the relative weighting of the dimensions may be partially determined by the method chosen. However, it is too early to conclude...
that this is the case. For instance, the observed differences in dimension importance could be caused by respondents being more influenced by the ordering of dimensions when ranking than when performing TTO. The order in which dimensions are presented is clearly construct-irrelevant. Fortunately, if this is the case, the observed differences should disappear if the ordering of the five dimensions was randomized. Alternatively, it could be that time framing is more salient in TTO and that respondents find the thought of long-term impairments on pain/discomfort or anxiety/depression unbearable. This interpretation is compatible with the repeated findings of non-linear time preferences and the concept of maximum endurable time in TTO\textsuperscript{171,190–192}, and with the findings in a cognitive debriefing study of EQ-5D valuation by Bailey and Kind, where respondents frequently ignored the 10 year duration in the VAS and ranking tasks, but were sensitive to time when performing TTO\textsuperscript{163}.

Some researchers argue that, lacking a gold standard by which to compare health state values can be compared, agreement with rank values may be used to assess other valuation methods\textsuperscript{90,159}. In that light, comparison between TTO and rank values is non-trivial. As long as the reasons behind the observed differences remain unknown, the apparent inconsistencies between rankings by the two methods reduce the certainty with which we can interpret the relative weights attributed to the different EQ-5D dimensions as representing the perceptions of the public valuing the states.

5.2.7 TARIFTS AS MEASURES OF THE PREFERENCES OF THE RESEARCHERS

The legitimacy of using population-based tariffs for instruments like the EQ-5D stems from the implicit consent of the population in question\textsuperscript{193}. If the tariff is an adequate model of how the population values health, decisions regarding distribution of health care resources based on the tariff should be in accordance with the preferences of the people. For instance, the authors of the Argentinian
EQ-5D tariff state that tariffs represent “… votes for and against potential health states, and their tallies inform policymakers which health states the public prefers to others”\textsuperscript{41}. In countries without a national EQ-5D tariff, it is recommended that QALYs should be calculated using tariffs from countries of geographical proximity\textsuperscript{42} or that have similar cultural characteristics\textsuperscript{194}. However, our studies indicate that choices researchers make in designing valuation studies impact substantially on the resulting tariffs. In the case of WTD transformation and the US and UK tariffs, the effect was found to be greater in magnitude than differences attributable to cultural differences.

Consider the case of a hypothetical country situated in the Atlantic Ocean. Lacking a national EQ-5D tariff, the population of said country could choose between the US and the UK tariffs. This choice, which ideally should be based on cultural similarity, would in effect select a value set more influenced by the differences in methodological choices used in the two countries than by cultural differences. Arguably, the choice of tariff would be better informed if founded on considerations of the appropriateness of methodological choices underlying the candidate tariffs.

In addition to transformation methods, several other methodological choices influence values. For instance, altering the starting point of the task could be used to manipulate resulting tariff values. With a growing body of observed breaches of procedural invariance and proposed alternative methods, investigators are left with an increasing number of difficult choices, all of which influence resulting tariffs. Thus, in a sense, final tariff values are increasingly functions of the choices of the investigators. Obviously, the actual influence of methods is insensitive to the knowledge of their effects. However, investigators’ knowledge of exactly how methodological choices influence values leave them with greater power to intentionally manipulate tariffs in such a way as to achieve values in accordance with their own preconceptions. This ability of investigators to constrain elicited or
modeled tariff values to a range of preconceived acceptable values challenges the interpretation of tariff values as being primarily population-based.

To confound the issue, the custom of presenting and using tariff values with uncertainty estimates in the magnitude of <.01, based solely on the statistical precision of population mean estimates based on any single chosen method (‘method’ including all necessary steps, like specific TTO variant, WTD transformation, regression model etc.) effectively obscures the actual uncertainty of the absolute values in question. These values, with their implied consent from the general public, are then used as if they were true in a very absolute sense, in that most users do not question them.

5.2.8 EXPERIENCE OR PREDICTION

With the exception of paper V, the focus of each of the papers included in this thesis have been trained on aspects related to general population values for EQ-5D health states. Paper V shows how values for EQ-5D health dimensions elicited from the general population diverge systematically from how impairments on the same EQ-5D dimensions influence self-rated HRQoL.

As mentioned in the background chapter, there is a conceptual distinction between experienced (or hedonic) utility as described by Jeremy Bentham and decision utility, as described by Fisher. The two are similar, and the choice between them relatively unchallenging when they converge, i.e., when people want what they would enjoy. In his seminal work “Bounded rationality, ambiguity, and the engineering of choice”, March argues that the ability to predict future experienced utility is a requirement of rational decision making. Kahneman and Dolan argue in favor of measuring experienced utility directly, due to common failure of decision utility to predict future experienced utility.

Our findings imply that general population values perform badly as predictive utility measures, if our comparison of the two sources of values is seen as a test
of the ability of the general public to predict how people in impaired health
would report their general health as a function of dimension of impairment, the
general population performs at below chance level. The reasons for this
discrepancy are not well known. If the observed deviation stems from the general
population’s lack of knowledge of the real-world impacts of different kinds of
impaired health, the use of population values are at risk of measuring construct
irrelevant factors like misplaced fear. Consider a real-market situation like the
process of buying a car. Most potential car-buyers do their best to amass
information on their alternatives before deciding, including, for instance,
information on the experiences of people who own different kinds of cars. This
contrasts with valuation studies, in which respondents are asked to value health
states without gathering further information. This raises several questions: How
would the general population value health given access to further information on
health impairment? If informed of the observed differences between general
population and patient values, which of the two types of values would the
general public prefer to base health-care expenditures on? These are empirical
questions the answers to which could be highly relevant for this debate.

In this thesis, we have focused on problems with general population values. This
is not to say that patient values are problem-free. For instance, people adjust to
living with impaired health, but such adjustment is likely to vary considerably by
types of impairment, personality, life situation, social support, culture, etc.
Furthermore, health problems are not randomly distributed along axes such as
age, geography, or personality, further complicating measurement of experienced
health impairment. Problems related to the valuations of experienced health
must necessarily be weighed against problems with valuing hypothetical health.
6 IMPLICATIONS AND FUTURE PERSPECTIVES

The implications of our findings are not limited to research. One of the most important issues raised in this thesis is QALY’s lack of delineation with regard to death. Both research on and practical use of the QALY would gain from a more well-defined ‘death’, with clear implications for what death-related aspects, if any, should be considered relevant. The rationale for, and the evidence supporting, QALY weights equal to or below death should be carefully reconsidered, since restricting the QALY to the BTD scale would avoid many problems, both theoretical and practical. There is a call for research into what respondents include in their valuation of health compared to ‘death’, and into ways of guiding what aspects they should consider.

Future research on TTO and other valuation methods should consider biases from the communicative implicatures of the valuation tasks, in addition to susceptibility to heuristics typically considered in behavioral economics.

As things stand, there is reason to caution users of EQ-5D tariffs against using them unquestioningly; in particular, the absolute values presented in national tariffs should be interpreted with care. The actual uncertainty surrounding these values is substantial (e.g., as reflected in differences between transformation methods, the observed anchoring effects, and the influence of ATE) but essentially unknown. In a health care resource allocation scheme where all interventions are compared using the same tariff, this uncertainty with regard to the absolute level of tariff values could be said to influence all QALY estimates equally, given that no patients move between impaired health and either ‘death’ or ‘perfect health’.

In our papers and this thesis, focus has been trained primarily on TTO, and to a limited extent VAS, in the context of EQ-5D valuation. The findings should not be interpreted as criticizing EQ-5D per se; other MAUIs using similar means of
valuation are likely to be similarly influenced. We have not considered SG directly, but the shared features of SG and TTO (starting point, routing procedure, comparison to ‘death’) suggest that SG is likely to be influenced by some of the issues observed for TTO.

Future valuation studies, should note that the procedure used to elicit health state values may impact on the resulting values. We have only investigated a few of many potentially influential factors. We stress the importance of considering which factors are construct-irrelevant and the importance of empirically testing whether these factors affect the results. Identifying factors that could influence the different EQ-5D dimensions differently is of particular importance.

We also call for research into how respondents from the general public would value health, given more information on how health impairments are experienced, and into whether or not they would prefer to use population-based or patient-based values for resource allocation purposes.
7 CONCLUSIONS

TTO-based tariff values appear to be substantially influenced by construct-irrelevant factors, including choice of WTD transformation method, anchoring from the starting point, unintended communicative implications, and fear of death. Additionally, factors of disputable relevance also influence TTO values, such as attitudes toward euthanasia. Combined, these factors make the absolute scale of TTO-values highly uncertain, and their interpretation in relation to the QALY-anchors of ‘perfect health’ and ‘death’ questionable.

It remains uncertain whether it is possible to validly measure health states equal to or worse than death.

Several practical and theoretical issues pertinent to QALY weights are related to values at or below the value of death. Restricting the QALY system to the >0 range should be reconsidered.

Given that ranking is considered by some to be the best available gold standard for assessing the consistency of other valuation methods, the observed inconsistency between rank and TTO values in terms of relative importance for the different EQ-5D dimensions is worrying.

Health state values based on general population respondents fail to predict what dimensions of health lead to reduced experienced HRQoL. This underlines the importance of the debate regarding what group’s values should be used in tariff generation.
1. Due to technical problems in the generation of the final PDF, the numbering of all level three headings was disconnected from the numbering of the other levels of headings. For example, section 1.2.1 (which is the first subsection of section 1.2) was numbered as section 1.1.1. The next was numbered as section 1.1.2. This continued into the following chapters, rendering the final level-three heading, which should be 5.2.8, to be numbered as 1.1.36. The numbering has been corrected to comply with the one presented in the index.

2. Page breaks have been altered to avoid tables being split.

3. The phrase “attitudes towards euthanasia” has been changed to “attitudes toward euthanasia” throughout.

4. Preface, first paragraph, second sentence:
   “This may be true for most PhD project, but probably not to this extent.”
   The word “project” has been changed to “projects”.

5. Section 1.3.2.2 – Conventional methodology and existing national tariffs.
   The word “methodology” has been changed to “methods”. The section is not concerned with the epistemological study of methods.

6. Section 1.4.2 – MVH-style TTO (listed as section 1.1.6), first paragraph, first sentence:
   “In MVH-style TTO, respondents may describe presented health states as being better than death (BTD), equal to death (ETD) or worse than death (WTD).
   A comma has been added after “(ETD)”.

7. Section 3.2.2 – The structure of the surveys, first paragraph, second sentence:
   “The first of each interview [...]”
   Changed to:
   “The first part of each interview [...]”

8. Section 3.4 – Permissions and ethics, paragraph five, first sentence:
   “The valuation experiments on which the studies presented in papers VI and VII were based were performed under the umbrella [...]”
   The sentence has been simplified for clarity. Now reads:
   “The valuation experiments presented in papers VI and VII were performed under the umbrella [...]”
9. Section 3.4, paragraph six first sentence:
   “Since no ethical research committee considered study to be inside their jurisdiction [...]”
   The word “the” has been added between “considered” and “study”.

10. Section 3.4, second paragraph, 5th sentence:
   “Our surveys involved TTO, which may discomfort respondents with its direct comparisons of impaired health lives to “immediate death”.
   The word “lives” has been removed.

11. Section 5.1.4.3 Paper III, third paragraph, second sentence:
   “We initially performed separate analyses on each on the 42 [...]”
   The second “on” has been changed to “of”.

12. Section 5.1.4.5 – Paper V, paragraph four, first sentence:
   “To render EH and HH values more comparable, we devised a set of procedures that allowed us focus solely on [...]”
   The word “to” has been added between “us” and “focus”.

13. Page 92, section 5.1.4.5 – Paper V, paragraph six, second sentence:
   “A can be seen from figure 10 [...]”
   The word “A” has been changed to “As”

14. Section 5.2.3 – The initial choice task, third paragraph, final sentence:
   “Consider, if only anchoring was at work, one would expect the tendency to elicit any specific to be a positive function of proximity with the starting point.”
   The word “value” has been added between “specific” and “to”.

15. References 65 and 185 were identical, and have been merged to 65.
9 REFERENCES


76. Linden WJ van der, Hambleton RK. *Handbook of modern item response theory.* Springer; 1997.


140. Costa PT, McCrae RR, Psychological Assessment Resources I. Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo- ffi). Psychological Assessment Resources; 1992.


10 Appendix: Survey Questions

For reference, we have included a copy of all the non-valuation questions asked of respondents in our surveys. The included copy is taken from the postal part of a 15D survey performed prior to our EQ-5D valuation experiments, and was chosen because it is better suited for print than screenshots. The relevant parts of the survey can be found over the next 6 pages.
1. Hvilket årstall er du født?
   19

2. Er du mann eller kvinne?
   Mann
   Kvinne

3. Hva er din høyeste fullførte utdanning?
   Grunnskole
   videregående utdanning (Allmennfag, yrkesskole eller annet)
   Fagsivilutdanning/yrkesutdanning/fagbrev/videregående yrkesfaglig
   Universitet/høgskoleutdanning med inntil 4 år
   universitet/høgskoleutdanning med mer enn 4 år

4. Hva er din sysselsetting?
   Inntektsgivende arbeid heltid
   Inntektsgivende arbeid deltid
   Selvstendig næringsdrivende
   Fritidsjobb
   Annen type jobb
   Elev, student
   Hjemmeværende/husarbeid i hjemmet
   Annet

5. Personinntekt?
   Under 200.000
   Mellom 200.000 - 299.999
   Mellom 300.000 - 399.999
   Mellom 400.000 - 599.999
   Mellom 600.000 - 799.999
   800.000 eller mer
   Ønsker ikke å svare

6. Hva er din sivilstatus?
   Gift/registrert partner
   Samboende
   Ugift
   Tidligere gift (separert/skilt/enke/enkemann)

7. Har du erfaring med alvorlig sykdom?
   Merk: Sett et kryss på hver linje.
   Ja
   Nei

8. Når det gjelder din egen helse:
   I egen familie:
   Pleie av syke/funksjonshemmede personer:

9. Har du en funksjonshemming?
   Merk: Sett et kryss.
   Ja
   Nei

10. Har du en kronisk sykdom diagnostisert av lege?
    Merk: Sett et kryss
    Ja
    Nei

11. Dersom ja på spørsmål 9, hvor mange kroniske sykdommer har du?
    Merk: Sett et kryss
    1
    2
    3
    4 eller flere

12. Synes du at helsetilstanden din i dag er redusert på grunn av sykdom eller funksjonshemming?
    Merk: Sett et kryss
    Ja
    Nei

13. Hvor er du født?
    Merk: Sett et kryss
    Norge
    Norden (utenom Norge)
    Europa (utsnitt av Norden)
    USA/Oceania
    Afrika/Asiak

14. Røyker du?
    Merk: Sett et kryss
    Nei, aldri
    Nei, har sluttet for mer enn 6 mnd siden
    Ja, av og til
    Ja, daglig under 15 sigaretter
    Ja, daglig over 15 sigaretter
Vi vil først be deg om å vurdere din egen helse på fem grunnleggende områder. Vennligst angi hvilke utsagn som best beskriver din helsetilstand i dag for hvert av de fem områdene under:

<table>
<thead>
<tr>
<th>15</th>
<th>Gange</th>
<th>Merk: Sett ett kryss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jeg har ingen problemer med å gå omkring</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeg har litt problemer med å gå omkring</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeg er sengeliggende</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>15</th>
<th>Personlig stell</th>
<th>Merk: Sett ett kryss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jeg har ingen problemer med personlig stell</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeg har litt problemer med å vaske meg eller kle meg</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeg er ute av stand til å vaske meg eller kle meg</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>16</th>
<th>Vanlige gjøremål (f.eks. arbeid, studier, husarbeid, familie- eller fritidsaktiviteter)</th>
<th>Merk: Sett ett kryss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jeg har ingen problemer med å utføre mine vanlige gjøremål</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeg har litt problemer med å utføre mine vanlige gjøremål</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeg er ute av stand til å utføre mine vanlige gjøremål</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>17</th>
<th>Smerte og ubehang</th>
<th>Merk: Sett ett kryss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jeg har verken smerte eller ubehang</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeg har moderat smerte eller ubehang</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeg har sterk smerte eller ubehang</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>18</th>
<th>Angst og depresjon</th>
<th>Merk: Sett ett kryss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jeg er verken engstelig eller deprimert</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeg er noe engstelig eller deprimert</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeg er svært engstelig eller deprimert</td>
<td></td>
</tr>
</tbody>
</table>
På en skala fra 0 til 100, hvor god eller dårlig vil du si at helsetilstanden din er i dag?

Vennligst svar ved å trekke en linje fra pilen nedenfor til det punktet på skalaen som viser hvor god eller dårlig din helsetilstand er i dag. Bruk gjerne en linjal. Notér deretter tallet som beskriver din helsetilstand i ruten under (0 – 100)

Tall mellom 0 og 100. Ett tall pr rute
Du er nå ferdig med spørsmålene som gjelder vurdering av helse. På de neste sidene har vi noen spørsmål om hvordan du vil beskrive deg selv. Det er ikke riktige eller gale svar på disse spørsmålene, og det du svarer vil ikke ha noen innvirkning på hvordan helsevesenet vil prioritere i fremtiden. Vi håper likevel du vil svare på spørsmålene, ettersom de er viktige for vår forskning.


<table>
<thead>
<tr>
<th>Svar</th>
<th>Svært uenig</th>
<th>Uenig</th>
<th>Verken enig eller uenig</th>
<th>Enig</th>
<th>Svært enig</th>
<th>Vet ikke</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeg er ikke en person som pleier å bekymre meg</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Til tider har jeg følt meg harm og forbittet</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Når jeg er svært stresset, føles det av og til som om jeg går i stykker</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Jeg føler meg sjelden ensom eller nedfor</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Jeg føler meg ofte anspent og nervøs</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Noen ganger føler jeg meg fullstendig verdsløs</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Jeg føler meg sjelden redd eller engstelig</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Jeg blir ofte sint over måten folk behandler meg på</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Når noe går galt, blir jeg altfor ofte motløs og får lyst til å gi opp</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Jeg er sjelden trist eller deprimert</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Jeg føler meg ofte hjelpeløs og ønsker at andre skal løse problemene mine</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Til tider har jeg vært så skamfull at jeg bare har ønsket å gjemme meg</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>
Spørsmålene på denne siden handler om dine holdninger og overbevisninger. Målet med disse spørsmålene er å studere sammenhengen mellom hvordan forskjellige mennesker verdsetter helsetilstander og hva de mener disse spørsmålene. Måten du svarer på disse spørsmålene vil ikke ha noen direkte konsekvenser for fremtidige prioriteringer i helsevesenet.

I hvilken grad vil du si at du er enig eller uenig i følgende utsagn:

### 57. Jeg anser meg selv som en religiøs person?
*Merk: Sett ett kryss*

- [ ] Svært uenig
- [ ] Uenig
- [ ] Verken enig eller uenig
- [ ] Enig
- [ ] Svært enig
- [ ] Vet ikke

### 58. Jeg tror på et liv etter døden.
*Merk: Sett ett kryss*

- [ ] Uenig
- [ ] Kanskje
- [ ] Enig
- [ ] Vet ikke
En person har en smertefull, uhelbredelig og dødelig sykdom. Livet nærmer seg slutten og han/hun har et veloverveid ønske om å dø, og ber om hjelp til dette.

<table>
<thead>
<tr>
<th>Hvor enig eller uenig er du i følgende utsagn?</th>
<th>Svært uenig</th>
<th>Uenig</th>
<th>Verken enig eller uenig</th>
<th>Enig</th>
<th>Svært enig</th>
<th>Vet ikke</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legen bør ha adgang til å avslutte livsfølgende behandling, som f. eks kunstig ernæring eller respiratorbehandling, slik at pasientens liv forkortes.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Legen bør ha adgang til å utføre en handling, f. eks gi en injeksjon, som bevisst tar sikte på å forkorte pasientens liv på en smertefri måte (aktiv dødshjelp).</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Legen bør ha adgang til å hjelpe personen slik at denne selv kan avslutte sitt liv på en smertefri måte (assistert selvmord).</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>