

Momentous Choices

Testing nonstandard decision models
in health and housing markets

Gewichtige keuzes

Het toetsen van niet-standaard beslissingsmodellen
in gezondheid en woningmarkten

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam

by command of the
Rector Magnificus
Prof.dr. H.G. Schmidt

and in accordance with the decision of the Doctorate Board

The public defense shall be held on
Friday 18 October 2013 at 11.30 hours

by

Martin Filko

born in Trstená, Slovakia



Doctoral Committee

Promoter

Prof.dr. H. Bleichrodt

Other members

Prof.dr. E.K.A. van Doorslaer

Dr. N.J.A. van Exel

Prof.dr. P.P. Wakker

Stellingen (Claims)

- 1) QALYs (Quality Adjusted Life Years) describe the preferences of a representative agent well. Previously observed violations were probably due to violations of expected utility rather than to violations of the QALY model itself.
- 2) Generalized statements about preferences are possible even if a sample is not representative and the number of observations is limited.
- 3) In many everyday situations, people do not optimize a classical utility function, but rather they decide based on similarities.
- 4) Geographical vicinity is more important than the type of dwelling (house versus apartment) in forming expectations about the price development of real estate.
- 5) Distributions are as important as averages. Estimating social welfare functions of individuals is therefore as crucial for applications of decision theory in practical policy making as estimating preferences of a representative agent.
- 6) A QALY is not a QALY. There are other important considerations at play.
- 7) Low numbers of transactions in real estate market downturns can be attributed to loss aversion and the endowment effect. Prospect theory can explain why it makes sense to use a lottery rather than a straightforward sale in a bear real estate market.
- 8) Revealed preference is an important principle in both scientific research and real life.
- 9) Most people, in most settings, exhibit other-regarding preferences.
- 10) Estate tax is a fair and efficient means of raising government revenue.
- 11) Polygraph is an unreliable means of establishing a person's general trustworthiness or specific truth telling. It is easy to cheat.

Contents

Introduction	7
Utility Independence of Multiattribute Utility Theory is Equivalent to Standard Sequence Invariance of Conjoint Measurement	12
2.1 Introduction	13
2.2 Notation	15
2.3 Utility independence	16
2.4 Standard sequence invariance	17
2.5 Generalizations and main result	20
2.6 An application to health	21
2.7 Conclusion	23
New Tests of QALYs when Health Varies over Time	35
3.1 Introduction	36
3.2 Background	39
3.3 Experiment	44
3.4 Results	51
3.5 Discussion	58
A Reply to Gandjour and Gafni	69
4.1 First criticism: Support for generalized marginality and violations of the QALY model can coexist	71
4.2 Second Criticism: No General Statements are Possible	74
4.3 Conclusion	76
Making Case-Based Decision Theory Directly Observable	78
5.1 Case-based decision theory versus classical revealed preference: varying memory instead of available choice options	81
5.2 Direct (nonparametric) measurements of utility and similarity weights: Theory	84
5.3 A CBDT version of the random incentive system	88
5.4 Experiment 1 to measure similarity weights and to test CBDT	89
5.4.1 Stimuli	89
5.4.2 Similarity weights in our design	96
5.4.3 Sample and procedure	97

5.4.4 Predictions of CBDT	99
5.4.5 Analysis	100
5.4.6 Results: Tests of CBDT	101
5.4.7 Results: Explorations regarding real estate investments	102
5.4.8 Discussion of Experiment 1	102
5.5 Experiment 2 to measure similarity weights and to test CBDT	103
5.5.1 Stimuli to measure similarity weights	103
5.5.2 Sample and procedure	105
5.5.3 Predictions of CBDT	105
5.5.4 Results: Tests of CBDT	106
5.5.5 Results: Explorations regarding real estate investments	106
5.5.6 Discussion of Experiment 2	106
5.5.7 A Comparison Between Experiment 2 and Experiment 1	107
5.6 General discussion	107
5.7. Summary and Conclusion	109
Conclusion	118

mo•men•tous [mō' mentəs]

Adjective

(of a decision, event, or change) Of great importance or significance, especially in its bearing on the future.

Synonyms

important – weighty – significant – grave – serious

Chapter 1

Introduction

During more than half a century, several strands of research contributed to the development of decision theory. The standard normative model for choice under uncertainty – *expected utility* – was given a foundation by von Neumann and Morgenstern (1944) and Savage (1954). It advised – and expected – reasonable actors to evaluate the consequences of their actions by the weighted sum of their utility, using probabilities of these consequences as weights. Utilities were derived from the choices made by actors themselves, and together with probabilities should be evaluated in a simple linear fashion. It can be shown that under certain reasonable conditions, behavior resulting from the theory is rational.

Economists traditionally assumed that the standard model is also valid descriptively (Arrow, 1951). They assumed that, even though individual human beings can err and deviate from the theory, these deviations are not systematic, and the theory's predictions for the consequences of economic action hold well (e. g., the *as-if hypothesis*, Friedman, 1953). However, with the advent of decision research in psychology, the building began to crack. Allais (1953) showed that reasonable people could make a series of choices violating expected utility. And, interestingly, stick with them even after the violations are pointed out. Even more fundamentally, Ellsberg (1961) showed that people may not be able to assign subjective probabilities to events.

Mathematical psychologists Kahneman and Tversky went beyond simple cataloging the quirks of human decisions. Their *prospect theory* (Kahneman and Tversky, 1979) was an extension of the standard model, simple enough to be used in both economic theorizing and interdisciplinary discussions, and powerful enough to explain some of the most striking paradoxes of human choice. It was also based on three intuitive notions – that people give more weight to small probabilities of extreme outcomes, that people evaluate consequences as changes from the reference point rather than as static states of the world, and that losses looms larger than gains. Importantly, it showed that Allais' and Ellsberg's findings were not an oddity for extreme choice

situations of minor importance, but were central in economics, affecting central questions there.

In the meantime, applications of decision research began to flourish. Medical decision making was one of the domains where applications of descriptively adequate decision models come naturally, given the importance – both economic and human – of the decisions being made. Limitations to using standard cost-benefit models based on life-expectancy as an outcome measure were partially overcome by the development of utility-based outcome indicators such as the QALY (Quality-Adjusted Life Years) model (Pliskin, Shepard and Weinstein, 1980). QALYs, a product of health quality and life duration if health is static, or the sum of health qualities over the specific time-points in general, turned out to be both intuitive enough for policy makers and tractable enough to be used in theory.

To explain actual choices that humans make, some researchers took a route different to tinkering with the standard states-of-nature model of Savage (1954). Gilboa and Schmeidler (1995) noticed that some decision situations lend themselves to thinking in analogies, rather than in terms of probabilities. In their *case-based decision theory* (CBDT), outcomes resulting from actions are still being evaluated in utility terms. Nonetheless, they are being weighted by their similarity to the previous situations rather than by their probabilities.

In the coming chapters of this thesis, I will draw heavily from the research streams described above. Some of the questions raised by a theoretical research in nonstandard decision making will be investigated empirically, especially in an applied context of health and real estate.

One of the more general questions in health decision making is whether life quality and life duration are separable. Utility independence is a central condition in multiattribute utility theory, where attributes of outcomes are aggregated in the context of risk. The aggregation of attributes in the absence of risk is studied in conjoint measurement. In conjoint measurement, standard sequences have been widely used to empirically measure and test utility functions, and to theoretically analyze them. Chapter 2 of this thesis shows that utility independence and standard sequences are closely related: Utility independence is equivalent to a standard

sequence invariance condition when applied to risk. This simple relation between two widely used conditions in adjacent fields of research is surprising and useful. It facilitates the testing of utility independence because standard sequences are flexible and can avoid cancellation biases that distort direct tests of utility independence. Extensions of our results to non-expected utility models such as prospect theory can now be provided easily. We discuss applications to the measurement of quality-adjusted life-years (QALY).

The QALY model has been mostly refuted for chronic health states (for an overview see Bleichrodt and Pinto-Prades, 2006), but what if this was caused by violations of the confounding assumption of expected utility rather than of the additive QALY model itself? Chapter 3 performs new tests of the QALY model when health varies over time. These tests do not involve confounding assumptions and are robust to violations of expected utility. The results support QALYs at the aggregate level, i.e. in economic evaluations of health care. At the individual level, there is less support for QALYs. The individual data are, however, largely consistent with a more general QALY-type model that remains tractable for applications.

The paper constituting Chapter 3 has been criticized by Gandjour and Gafni (2010) on two counts. First, they argue that it is possible that the condition tested, generalized marginality, is not sufficient to imply the QALY model. In other words, subjects may simultaneously satisfy generalized marginality and violate the QALY model. Second, Gandjour and Gafni argue that we cannot make generalized statements about preferences because our sample is not representative. Related to this, they argue that we cannot conclude in support of a particular model based on a limited number of tests because the variety of health profiles is essentially endless. In Chapter 4, I show that Gandjour and Gafni's first point of criticism is wrong. Their arguments contain many mathematical mistakes implying that their counterexamples are wrong and, therefore, that all their corresponding speculations are irrelevant. Their other points of criticism are completely standard (Popper, 1934, 1963) and reflect a lack of understanding of the general principles underlying all empirical studies in all fields of science. Moreover, these points have actually been acknowledged and discussed in our original paper as they are in Chapter 4.

In Chapter 5, I turn my attention to an alternative explanation of human decision making. What if people do not optimize a classical utility function, but rather decide based on similarities? The latter is a natural approach in many circumstances. If we choose a dish from a menu then we think of similar experiences in the past. Gilboa & Schmeidler's case-based decision theory (CBDT) is an alternative to Savage's state-space model for uncertainty. Preferences are determined by similarities with cases in memory. A difficulty in experimental implementations of CBDT has so far been that not only the income effect, well known from classical theories, but also interaction between different memories assumed in different experimental questions, has to be avoided, and no way was known hitherto to avoid such interactions. Chapter 5 introduces such a method to elicit CBDT, requiring no commitment to parametric families and relating directly to decisions. An experiment on real estate investments demonstrates the feasibility of the method. I confirm CBDT's predictions with however one violation, being separability of cases in memory. I conclude that CBDT gives plausible predictions and new insights into (real estate investment) decisions.

References

- Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica* 21 (4), 503–546.
- Arrow, K. J. (1951). Alternative Approaches to the Theory of Choice in Risk-Taking Situations. *Econometrica* 19, 404–437.
- Bleichrodt, H., Pinto-Prades, J. L. (2006). Conceptual foundations for health utility measurement. In: Jones, A.M. (Ed.), "The Elgar Companion to Health Economics. Edward Elgar" Aldershot, 347-358.
- Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics*, Vol. 75, No. 4 (Nov., 1961), 643-669.
- Gilboa, I., Schmeidler, D. Case-Based Decision Theory. *The Quarterly Journal of Economics* (1995) 110 (3): 605-639.

Kahneman, D., Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*.

Milton, F. (1953). *Essays in Positive Economics*. Chicago University Press.

von Neumann, J. & Morgenstern, O. (1944, 1947, 1953). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Pliskin, J. S., Shepard, D. S., Weinstein, M. C. (1980). Utility functions for life years and health status. *Operations research*.

Popper, K. (1934). *Logik der Forschung*.

Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. Routledge & Kegan Paul. London.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley. Second Edition, New York: Dover, 1972.

**Utility Independence of Multiattribute Utility Theory
is Equivalent to Standard Sequence Invariance
of Conjoint Measurement**

Summary

Utility independence is a central condition in multiattribute utility theory, where attributes of outcomes are aggregated in the context of risk. The aggregation of attributes in the absence of risk is studied in conjoint measurement. In conjoint measurement, standard sequences have been widely used to empirically measure and test utility functions, and to theoretically analyze them. This chapter shows that utility independence and standard sequences are closely related: Utility independence is equivalent to a standard sequence invariance condition when applied to risk. This simple relation between two widely used conditions in adjacent fields of research is surprising and useful. It facilitates the testing of utility independence because standard sequences are flexible and can avoid cancellation biases that affect direct tests of utility independence. Extensions of our results to nonexpected utility models can now be provided easily. We discuss applications to the measurement of quality-adjusted life-years (QALY) in the health domain.

Keywords: Utility independence, standard sequences, multiattribute utility, conjoint measurement, nonexpected utility.

¹ This chapter was published as Bleichrodt, H., Doctor, J. N., Filko, M., Wakker, P. P. (2011). Utility Independence of Multiattribute Utility Theory is Equivalent to Standard Sequence Invariance of Conjoint Measurement. Han Bleichrodt's research was made possible by a grant from the Netherlands Organisation for Scientific Research (NWO). Martin Filko's research was made possible by a grant from DSW health insurance.

2.1 Introduction

Utility independence is widely used in decision analysis (Keeney & Raiffa, 1976; Guerrero & Herrero, 2005; Engel & Wellman, 2010). In medical decision making, utility independence underlies the health utility index, a widely used method to derive utilities for multiattribute health states (Feeny, Furlong, Torrance, Goldsmith, Zhu, Depauw, Denton, & Boyle, 2002; Feeny, 2006). Analyses of utility independence are usually based on the normatively convincing, but descriptively problematic, expected utility theory for choices between risky prospects (probability distributions over outcomes). Then the condition usually implies that multiattribute utility is additive, multiplicative, or multilinear.

Utility independence concerns situations where the levels of some attributes are fixed deterministically. The condition then requires that preferences between prospects over the remaining attributes should be independent of the fixed deterministic levels. This requirement has often been tested directly (Miyamoto & Eraker, 1988; Bleichrodt & Johannesson, 1997; Bleichrodt & Pinto, 2005; Spencer & Robinson, 2007). One problem with direct tests of utility independence is that they induce subjects to ignore the common fixed values, not because this is their true preference but rather as a heuristic to simplify the task before any consideration of true preference (Kahneman & Tversky, 1979, the *cancellation heuristic*). That such distorting heuristics can sometimes increase consistency, misleadingly suggesting verification of preference conditions, was emphasized by Loomes, Starmer, & Sugden (2003). For direct tests of utility independence the cancellation heuristic will indeed create artificial support for the condition.

A second problem with traditional analyses of utility independence is that they have been based on expected utility maximization. There is, however, much evidence that expected utility is violated empirically (Allais, 1953; Ellsberg, 1961; Kahneman and Tversky 1979; Starmer, 2000). Extensions of utility independence to nonexpected utility models include Bier & Connell (1994), Bleichrodt, Schmidt, & Zank (2009), Bouyssou & Pirlot (2003), Dyckerhoff (1994), and Miyamoto & Wakker (1996).

The aggregation of attributes is also studied in conjoint measurement (Krantz, Luce, Suppes, & Tversky, 1971). Unlike multiattribute utility theory and decision analysis, conjoint measurement does not assume risk to be present. However, one can still use the techniques of conjoint measurement in the presence of risk. This is the approach to multiattribute utility taken in this chapter. A common technique underlying many results in conjoint measurement is the construction of standard sequences.² These are sequences of attribute levels that are equally spaced in utility units, endogenously derived from preferences without using the utility function. In marketing, standard sequences are used in the saw-tooth method (Fishburn, 1967; Louviere, Hensher, & Swait, 2000). Krantz et al. (1971) explain the importance of standard sequences in great detail. Many preference conditions amount to invariance of particular standard sequences. By imposing such specific invariance conditions, specific functional forms of the multiattribute utility function can be derived.³

This chapter shows that there exists a surprisingly simple relation between multiattribute utility and conjoint measurement: utility independence is equivalent to a version of standard sequence invariance. This opens new and useful ways to analyze utility independence. Standard sequence techniques are flexible and efficient and they can avoid the aforementioned cancellation bias. Further, they give direct quantitative measurements of utility, which is useful in its own right. They do not directly appeal to risk, as does utility independence, but they focus on tradeoffs between attributes, avoiding the complications of risky decisions. Finally, they can easily be extended to nonexpected utility models, offering the possibility to design tests of utility independence that are robust to violations of expected utility.

² See Abdellaoui (2000), Baron (2008, Chs. 10 and 14), Booij & van de Kuilen (2009), Fishburn & Rubinstein (1982, pp. 682-3 and Figure 1), Loewenton & Luce (1966), von Winterfeldt & Edwards (1986, p. 267).

³ See Bouyssou & Pirlot (2004), Ebert (2004), Fishburn & Edwards (1997, Axiom 8), Gilboa, Schmeidler, & Wakker (2002) Harvey (1986, p. 1126), Casadesus-Masanell, Klibanoff, & Ozdenoren (2000), Krantz et al. (1971), Nau (2006, Axiom 4), Schmidt (2003), Skiadas (1997), Stigler (1950), Tversky & Kahneman (1992), Tversky, Sattath, & Slovic (1988), Wakker (1984), Wakker (2010), Wakker & Tversky (1993).

2.2 Notation

We start by assuming a simple model on a simple domain (a rank-ordered set of binary prospects) that is present as a substructure in expected utility but also in most nonexpected utility models. In all these models, the theorems that we obtain within the simple model immediately extend to the whole model. Consequently, our main result, Observation 5.2, applies to all these (non)expected utility models. Miyamoto and Wakker (1996) similarly used rank-ordered binary prospects to obtain results for many nonexpected utility theories.

We consider decision under uncertainty with one *event* E . E is uncertain in the sense that the decision maker does not know for sure if it is true (“will happen”) or not. An objective probability p of E may (the case of risk) or may not (the case of uncertainty and ambiguity) be given. Our analysis applies to either case. We consider *prospects* $x_E y$ yielding *outcome* x if E is true and outcome y otherwise. If an objective probability p is given for E , then we can also write $x_p y$. X denotes the *outcome set*.

A preference relation \succsim is given over the outcomes. The domain of prospects is *rank-ordered*: We assume without further mention that always $x \succsim y$ in prospects $x_E y$. The resulting rank-ordered⁴ set of prospects is denoted X_{\downarrow}^2 . A preference relation \succsim' is given on X_{\downarrow}^2 . *Constant prospects*, $x_E x$, yielding outcome x for sure are identified with that outcome x . The preference relation \succsim' generated over outcomes is assumed to agree with \succsim . Thus \succsim' defined over prospects is an extension of \succsim defined over outcomes. We will therefore write \succsim instead of \succsim' henceforth. Strict preference and indifference are defined as usual, and are denoted $>$ and \sim .

We assume that the outcome set X is a *two-attribute* product set $\mathcal{Q} \times \mathcal{T}$, with generic element $x = (Q, T)$. \mathcal{Q} designates the first attribute and \mathcal{T} designates the second, and \mathcal{Q} and \mathcal{T} are *attribute sets*. For example, if outcomes are chronic health states then \mathcal{Q} designates a health state and \mathcal{T} designates a time period (life duration). The extension of our results to cases of more than two attributes will be presented in §5.

⁴ Another widely used term in the literature is comonotonic.

We assume throughout that preferences over prospects $(Q_1, T_1) \succsim_E (Q_2, T_2)$ can be represented by

$$\pi U(Q_1, T_1) + (1-\pi)U(Q_2, T_2). \quad (2.1)$$

Here $U: \mathcal{Q} \times \mathcal{T} \rightarrow \mathbb{R}$ is the *utility function*, whose particular form is the central topic of multiattribute utility and of this chapter. The *decision weight* of event E is $0 < \pi < 1$. Equation (2) includes virtually all decision theories known today. Well-known examples are: (a) Expected utility where $\pi = P(E)$ is the probability of event E , objective in the case of risk and subjective in the case of uncertainty; (b) rank-dependent utility for risk (Quiggin, 1982) where $\pi = w(p)$ with p the objective probability of event E and w a probability weighting function; (c) rank-dependent utility for uncertainty (also called Choquet expected utility) or prospect theory where $\pi = W(E)$ with W a nonadditive weighting function or capacity (for gains under prospect theory); (d) maxmin expected utility (Gilboa & Schmeidler, 1989). Further details are in the footnote to Observation 5.2, and in Wakker (2010, §§6.11 and 10.6).

2.3 Utility independence

The second attribute \mathcal{T} is *utility independent* if

$$\begin{aligned} (Q, T_1) \succsim_E (Q, T_2) &\succsim (Q, T_3) \succsim_E (Q, T_4) \\ &\Leftrightarrow \\ (Q', T_1) \succsim_E (Q', T_2) &\succsim (Q', T_3) \succsim_E (Q', T_4) \end{aligned} \quad (3.1)$$

for all Q, Q' and for all T_1, T_2, T_3, T_4 . That is, preferences do not depend on the particular deterministic level at which Q is fixed. As throughout, it is implicitly assumed that all prospects are contained in X_{\downarrow}^2 . *Preferential independence* is utility independence restricted to constant prospects:

$$\begin{aligned} (Q, T_1) &\succsim (Q, T_3) \\ &\Leftrightarrow \\ (Q', T_1) &\succsim (Q', T_3). \end{aligned} \quad (3.2)$$

In economic consumer theory, preferential independence is known as separability of \mathcal{T} , and in conjoint measurement (Krantz et al., 1971) it is part of joint independence. Preferential independence implies that we can define preferences over the second attribute \mathcal{T} independently from the first attribute. It is naturally satisfied if \mathcal{T} is an interval and monotonicity holds. A convenient implication of preferential independence is that changing Q in Eq. 3.1 does not affect rank-ordering. That is, the upper two prospects in Eq. 3.1 are contained in X_{\downarrow}^2 if and only if the lower two are.

Utility independence of \mathcal{T} holds if U is *additive* ($U(Q,T) = V(Q) + W(T)$) or *multiplicative* ($U(Q,T) = V(Q)W(T)$) with all values $V(Q)$ of the same sign, which can then be taken positive. Under additional conditions, utility independence is not only necessary, but also sufficient for U being additive or multiplicative (Miyamoto and Wakker, 1996, Theorem 3). Then, in Eq. 3.3 below, f or g has to be constant. The following theorem extends a well known result from classical setups to our domain X_{\downarrow}^2 .

THEOREM 3.1. Assume that the image of the function $T \mapsto U(Q,T)$ is an interval for all Q . Then \mathcal{T} is utility independent if and only if

$$U(Q,T) = f(Q)V(T) + g(Q) \quad (3.3)$$

for some functions f, V, g with f positive. \square

2.4 Standard sequence invariance

A convenient feature of the standard sequence technique introduced next is that it is directly related to the empirical measurement of utility. T_0, \dots, T_n is a (Q -)standard sequence if there exist Q^*, T_g , and T_G such that, for $i = 0, \dots, n-1$,

$$(Q^*, T_g)_E (Q, T_{i+1}) \sim (Q^*, T_G)_E (Q, T_i) . \quad (4.1)$$

(Q^*, T_g) and (Q^*, T_G) are called *gauge outcomes*. They serve as a measuring rod to peg out the standard sequence. For later purposes, it is of interest to note that Q^* and Q

can be different. The proof of the following lemma is given in the main text because it may be clarifying.

LEMMA 4.1. Under Eq. 2.1, a Q-standard sequence is equally spaced in utility units ($U(Q, T_{i+1}) - U(Q, T_i)$ is independent of i).

PROOF. By Eq. 2.1, the $(1-\pi)$ weighted differences $U(Q, T_{i+1}) - U(Q, T_i)$ all match exactly the same π weighted difference $U(Q^*, T_g) - U(Q^*, T_g)$. \square

We now turn to comparisons of standard sequences for different values of Q . A Q-standard sequence T_0, T_1, T_2, \dots and a Q' -standard sequence T_0', T_1', T_2', \dots are *inconsistent* if they satisfy $T_0 = T_0'$ and $T_1 = T_1'$, but, for some $i > 1$, T_i and T_i' are not equivalent in the sense that $(Q, T_i) \not\sim (Q, T_i')$ or $(Q', T_i) \not\sim (Q', T_i')$.⁵ Under Eq. 2.1, inconsistencies are possible because equal spacedness for $U(Q, \cdot)$ need not correspond with equal spacedness for $U(Q', \cdot)$. *Standard sequence invariance on \mathcal{T}* means that such inconsistencies are excluded for all $Q, Q' \in \mathcal{Q}$.

THEOREM 4.2. Assume Eq. 2.1, with the image of the function $T \mapsto U(Q, T)$ an interval for each Q . Preferential independence of \mathcal{T} and standard sequence invariance on \mathcal{T} hold if and only if

$$U(Q, T) = f(Q)V(T) + g(Q) \tag{4.2}$$

for some functions f, V, g with f positive. \square

The comparison of Theorems 3.1 and 4.2 establishes an interesting connection between conjoint measurement and multiattribute utility because the necessary and sufficient form in Eq. 3.3 is identical to that in Eq. 4.2: Under preferential independence and richness, standard sequence invariance on \mathcal{T} is equivalent to utility independence of \mathcal{T} ! That is, we can test utility independence by testing standard sequence invariance. We can now for instance reduce the cancellation heuristic by taking different Q and Q^* in Eq. 4.1. This way, we can avoid biases that have distorted traditional tests of utility independence. We will state the relations between

⁵ It can be seen that Eq. 2.1 implies $Q' \neq Q$.

utility independence and standard sequence invariance formally in the following section.

We next provide an axiomatization of multiplicative utility, useful for QALY measurement in health (§6). We call $T_0 \in \mathcal{T}$ a *null element* if $(R, T_0) \sim (R', T_0)$ for all R and R' .

OBSERVATION 4.3. Assume that Eqs. 2.1 and 4.2 hold. If \mathcal{T} contains a null element then $g(Q)$ is constant and can be taken equal to 0, giving a multiplicative representation

$$U(Q, T) = f(Q)V(T). \quad (4.3)$$

□

For similar results, see Miyamoto, Wakker, Bleichrodt, & Peters (1998, Theorem 3.1) and Bleichrodt and Pinto (2005, Theorem 2). A remarkable implication of the above result is that \mathcal{Q} then also is utility independent on the subdomain where V is positive (which excludes the null element).

We have defined standard sequences for outcomes under not-E, that is, outcomes ranked worst and less preferred than the gauge outcomes. Standard sequences can equally well be defined for outcomes under E, when they are ranked best and are preferred to the gauge outcomes, using the following indifferences:

$$(Q, T_{i+1})_E (Q^*, T_g) \sim (Q, T_i)_E (Q^*, T_G). \quad (4.4)$$

For representation theorems, the topic of this chapter, it is desirable to use weak preference conditions in order to obtain the logically strongest theorems. For empirical investigations it can be interesting to consider more restrictive preference conditions, to obtain more possibilities to falsify a theory or to measure its concepts. Hence, for empirical purposes it may be interesting to also consider standard sequences defined in Eq. 4.4 and to investigate consistency properties between such larger classes of standard sequences. It easily follows that we should also have invariance here under Eq. 4.2.

Remark A.2 will indicate a mathematical generalization of our theorems that we do not present in the main text because it loses the empirically attractive reduction of the cancellation heuristic. An interesting feature of the weaker preference condition used there is that it is a common weakening of utility independence and standard sequence invariance. Thus the two conditions are different strengthenings of a common underlying necessary and sufficient condition. This observation clarifies the mathematical nature of our results.

2.5 Generalizations and main result

We first extend our results to n -attribute utility. Assume that X is $X_1 \times \cdots \times X_n$ for a natural number $n \geq 2$, with generic element (x_1, \dots, x_n) . Let $I \subset \{1, \dots, n\}$ and write $\mathcal{T} = \prod_{i \in I} X_i$ and $\mathcal{Q} = \prod_{i \notin I} X_i$. We can write $X = \mathcal{Q} \times \mathcal{T}$. *Utility independence* of I is defined as utility independence of \mathcal{T} (Eq. 3.1). That is, if the attribute levels outside of I are kept fixed at deterministic levels, then the preferences generated over prospects over \mathcal{T} are independent of the deterministic levels chosen. We can define standard sequences on $\prod_{i \in I} X_i$ exactly as in Eq. 4.1, where now $T_g, T_{i+1}, T_G, T_i \in \prod_{j \in I} X_j$, and $Q^*, Q \in \prod_{i \notin I} X_i$. Standard sequence invariance on $\prod_{i \in I} X_i$ requires consistency between standard sequences in $\prod_{i \in I} X_i$ for all Q and Q' in $\prod_{i \notin I} X_i$. The following theorem immediately follows from Theorems 3.1 and 4.2.

THEOREM 5.1. Assume a preference \succsim on X_{\downarrow}^2 , with $X = X_1 \times \cdots \times X_n$, and $I \subset \{1, \dots, n\}$. Let $\mathcal{T} = \prod_{i \in I} X_i$ and $\mathcal{Q} = \prod_{i \notin I} X_i$. Preferences are represented by Eq. 2.1 (with $T = (x_i)_{i \in I}$ and $Q = (x_i)_{i \notin I}$). The image of $(x_i)_{i \in I} \mapsto U((x_j)_{j \notin I}, (x_i)_{i \in I})$ is an interval for each $(x_j)_{j \notin I}$. Then I is utility independent if and only if $\prod_{i \in I} X_i$ is preferentially independent and standard sequence invariance on $\prod_{i \in I} X_i$ holds. \square

We next consider decision theories defined on general domains of prospects, leading to our main result. Now prospects can be probability distributions over outcomes with more than one probability involved, or mappings from multi-element state spaces to outcomes, and prospects need not all have the same rank-ordering. The definition of utility independence needs no adaptation: On all subproduct domains, preference is independent of the deterministic level at which outside attributes are kept fixed. We

define standard sequence invariance by defining standard sequences on all subsets isomorphic to X_{\downarrow}^2 (two outcomes and a fixed event or probability, always with the same rank ordering). No inconsistencies should result both within sets X_{\downarrow}^2 and across different sets X_{\downarrow}^2 . In many theories, this definition can be extended. For example, under rank-dependent utility it can be extended to all multi-event sets of prospects that are comonotonic (defined in Wakker 2010, §10.12). For brevity, we do not elaborate on this point.

OBSERVATION 5.2. Let $X = X_1 \times \dots \times X_n$ be a set of outcomes, and let \succsim be a preference relation on a set of prospects. Prospects can be probability distributions over X (risk), or functions from a state space S to X (uncertainty). The set of prospects is rich enough to contain a set of the form X_{\downarrow}^2 . Preferences are represented by a model that implies Eq. 2.1 on X_{\downarrow}^2 with the same utility function U as in Eq. 2.1 used throughout the domain. The utility function is an interval scale, i.e. preferences are not affected if a constant is added to utility or if utility is multiplied by a positive constant.⁶ If, for a set $I \subset \{1, \dots, n\}$, the utility image of $\prod_{i \in I} X_i$ is an interval whenever the attributes outside of I are kept fixed, then utility independence of I is equivalent to preferential independence and standard sequence invariance on $\prod_{i \in I} X_i$. \square

2.6 An application to health

This section applies the above results to medical decision making. Outcomes (Q, T) are chronic health states, with Q describing the constant health state and T the life

⁶ The requirements in our observation hold for most theories that are popular today. These include expected utility for risk (von Neumann & Morgenstern, 1944) and for uncertainty (Savage, 1954), rank-dependent utility for risk (Quiggin, 1982) and for uncertainty (Gilboa, 1987; Schmeidler, 1989), prospect theory if there are only gains (Luce & Fishburn, 1991; Tversky and Kahneman, 1992), disappointment aversion theory (Gul, 1991), maxmin expected utility (Gilboa and Schmeidler 1989; Wald, 1950) and the α -maxmin model (Hurwicz, 1951; Jaffray, 1994), contraction expected utility (Gajdos, Hayashi, Tallon, & Vergnaud, 2008), and binary rank-dependent utility (Luce, 2000, Ch. 3; Ghirardato & Marinacci, 2001; Wakker, 2010, §§6.11, 10.6). Observation 5.2 applies to all these theories.

duration spent in this health state, followed by death. Unlike in economics or psychology, statistical probabilities of risks are often available in the health domain. We will assume that prospects are probability distributions over chronic health states.

The utility of life duration T is described by a function V . The commonly found subjective time preferences and discounting imply that V is concave, with future life years contributing less to V than the first life years to come. Since the 1980s it has become customary to correct life duration for quality of life, leading to the QALY model $f(Q)V(T)$, where f designates the correction factor due to the subjective quality of life of health state Q . The QALY model is widely used in health policy.

Preference axiomatizations can serve to justify the use of QALYs as outcome measure (Pliskin, Shepard, & Weinstein, 1980; Miyamoto & Eraker, 1988; Bleichrodt & Quiggin, 1997; Bleichrodt, Wakker, & Johannesson, 1997; Miyamoto et al., 1998; Miyamoto, 1999; Bleichrodt & Miyamoto, 2003; Doctor & Miyamoto, 2003; Doctor, Bleichrodt, Miyamoto, Temkin, & Dikmen, 2004; Bleichrodt and Pinto, 2005). Observation 4.3, combined with Theorem 4.2, provides a new foundation of the QALY model with standard sequence invariance instead of utility independence. Here $T = 0$ life years naturally serves as the null element required by Observation 4.3. Standard sequence invariance entails that tradeoffs between life-years (discounting) are not different under different health states. This condition will sometimes be more intuitive than utility independence, which appeals to risk attitudes for life-years rather than to direct tradeoffs between life-years and intertemporal preferences.

Obviously, if standard sequence invariance is prescriptively objectionable then Observation 4.3 shows that the QALY model is prescriptively objectionable. Standard sequence invariance can also be used to test the descriptive (rather than prescriptive) validity of the QALY model. A tractable way of testing is as follows. First elicit a Q -standard sequence T_0, T_1, \dots, T_k through indifferences

$$(Q^*, T_g)_p (Q, \mathbf{T}_{i+1}) \sim (Q^*, T_G)_p (Q, T_i) .$$

as in Eq. 4.1, where the new value to be elicited in each indifference has been printed bold. Next take a health state $Q' \neq Q$ and a health state Q^{**} , which can be but need not be different from Q^* . Then use a “bridge” question

$$(Q^{**}, T'_g)_p (Q', T_1) \sim (Q^{**}, T_G)_p (Q', T_0)$$

to find new gauge outcomes $(Q^{**}, T'_g)^7$ and (Q^{**}, T_G) that should provide the same standard sequence starting with T_0 and T_1 . Then elicit a second standard sequence T'_0, T'_1, \dots, T'_k :

$$(Q^{**}, T'_g)_p (Q', T_{i+1}) \sim (Q^{**}, T_G)_p (Q', T'_i).$$

We can then test whether the two standard sequences agree, as required by standard sequence invariance and the QALY model. A useful spinoff of these measurements is that they directly measure the utility functions (i.e., discounting) for life duration under Q and Q' (Wakker & Deneffe, 1996). If these are different under Q than under Q' then the QALY model is violated.

The measurements proposed above are chained, with answers to one question serving as input of next questions. A drawback of chaining is that errors propagate. Our consistency questions indicated that the errors in most responses were modest. Simulation studies for standard sequences have suggested that the problem of error propagation is not very serious (Bleichrodt & Pinto, 2000, p. 1495; Abdellaoui, Vossman, & Weber, 2005, p. 1394, §5.3 end; Bleichrodt, Cillo, & Diecidue, 2010, p. 164; van de Kuilen & Wakker, 2011; Conte, Hey, & Moffatt, 2011).

2.7 Conclusion

We have demonstrated that standard sequences, a tool commonly used in conjoint measurement (where no risk is assumed), can also be used in multiattribute utility theory (where risk is assumed). They provide convenient tools to characterize and analyze utility independence, the most widely used preference condition in multiattribute utility theory. In particular, they facilitate the study of the QALY model for health decisions.

⁷ T'_g can but need not be equal to T_g .

Appendix. Proofs

PROOF OF THEOREM 3.1. That the functional form implies utility independence follows from substitution. Hence we assume utility independence, and derive the functional form.

Fix a Q^* . If the corresponding utility interval is one-point, then by utility independence preference is independent of T , V is constant, and everything follows. Hence, assume that the interval is nonpoint. Then with $V(T) = U(Q^*, T)$, this function is an interval scale in the representation $(T_1, T_2) \rightarrow \pi V(T_1) + (1-\pi)V(T_2)$, which means that it is unique up to level and unit. This uniqueness is well known if we have an expected utility representation on the full, nonrank-ordered, product set \mathcal{T}^2 (resulting from X^2 by keeping $Q = Q^*$ fixed), which is a special case of an additive conjoint representation with Krantz et al.'s (1971) restricted solvability satisfied.⁸ It is also well known if we have a rank-dependent representation on the full product set \mathcal{T}^2 (Wakker, 1991). That it also holds when restricted to the rank-ordered set \mathcal{T}_\downarrow^2 (resulting from X_\downarrow^2 by keeping $Q = Q^*$ fixed) as in our setup follows from Chateauneuf & Wakker (1993, Theorem 2.2 and Lemma C.4).

By utility independence the same preferences hold over pairs (T_1, T_2) with Q fixed at every other level $Q' \neq Q^*$. By interval scaling, we have $U(Q', T) = f(Q')V(T) + g(Q')$ with $f(Q')$ positive. This way we obtain the functions f and g . \square

PROOF OF THEOREM 4.2. If the functional form in the theorem holds, then all T s are ordered by V , implying preferential independence. Further, then all standard sequences are equally spaced in V units, and they must be consistent. This implies standard sequence invariance on \mathcal{T} .

In the rest of this proof we assume standard sequence invariance on \mathcal{T} and preferential independence and derive Eq. 4.2. By preferential independence we can define a

⁸ Here, and in what follows, we have continuity with respect to the product topology of the order topology generated over T , where the crucial point is that this topology is connected (it is also topologically separable). The result can be seen in more elementary terms if we transform all values T into $V(T)$, giving a weighted additive representation with linear value functions.

preference relation over \mathcal{T} independently of Q , that we will denote \succsim . Thus $T \succsim T'$ if $(Q, T) \succsim (Q, T')$ for some Q , which then holds for all Q .

Take some $Q \neq Q^*$. Define $V(T) = U(Q, T)$ and $V^*(T) = U(Q^*, T)$. By preferential independence, V and V^* both represent \succsim over \mathcal{T} and $V^* = \varphi \circ V$ for a strictly increasing φ that is continuous because it maps an interval onto an interval.

Take a T with $V(T)$ in the interior of $V(\mathcal{T})$. Hence, T is not maximal in \mathcal{T} . T will be fixed until the last lines in the proof. Define an open interval S around $V(T)$ so small that there is a “dominating” interval D in $V(\mathcal{T})$ above the interval S large enough to imply, for all T_1 and T_0 in $V^{-1}(S)$, existence of T_g and T_G in $V^{-1}(D)$ such that

$$(Q, T_g)_E (Q, T_1) \sim (Q, T_G)_E (Q, T_0). \quad (\text{A.1})$$

In words: each $(1-\pi)$ weighted V difference in S can be matched by a π -weighted V difference in D .

We similarly define an open interval S^* around $V^*(T)$ so small that there is a dominating interval D^* in $V^*(\mathcal{T})$ above the interval S^* large enough to imply, for all T_1 and T_0 in $V^{*-1}(S^*)$, existence of T_g^* and T_G^* in $V^{*-1}(D^*)$ such that

$$(Q^*, T_g^*)_E (Q^*, T_1) \sim (Q^*, T_G^*)_E (Q^*, T_0). \quad (\text{A.2})$$

That is, each $(1-\pi)$ weighted V^* difference in S^* can be matched by a π -weighted V^* difference in D^* .

Take a $T^+ > T$ so close to T that both $V(T^+) \in S$ and $V^*(T^+) \in S^*$. Similarly, take a $T^- < T$ so close to T that both $V(T^-) \in S$ and $V^*(T^-) \in S^*$. We consider the preference interval $\{T' \in \mathcal{T}: T^- < T' < T^+\}$ around T and two of its elements $T_0 < T_2$. We can find T_1 such that T_0, T_1 , and T_2 are equally spaced in V units, and T_1^* such that T_0, T_1^* and T_2 are equally spaced in V^* units.

LEMMA A.1. $T_1 \sim T_1^*$.

PROOF. (The end of the proof of this lemma will be indicated by *QED*.) For contradiction, assume $T_1 < T_1^*$ (the case with $>$ is similar and is not discussed).

Because the V values of T_0 , T_1 , and T_2 are contained in $V^{-1}(S)$, there exist T_g and T_G in $V^{-1}(D)$ such that, for $i = 0$:

$$(Q, T_g)_E (Q, T_{i+1}) \sim (Q, T_G)_E (Q, T_i). \quad (\text{A.3})$$

Because T_2 and T_1 have the same V difference as T_1 and T_0 , Eq. A.3 also holds for $i = 1$. That is, T_0, T_1, T_2 is a Q -standard sequence.

Because $T_1 < T_1^*$, we can find $T_2^* < T_2$ such that T_0, T_1, T_2^* are equally spaced in V^* units.

Similar to Eq. A.3, because the V^* values of T_0, T_1 , and T_2^* are contained in $V^{*-1}(S^*)$, there exist T_g^* and T_G^* in $V^{*-1}(D^*)$ such that

$$(Q^*, T_g^*)_E (Q^*, T_1) \sim (Q^*, T_G^*)_E (Q^*, T_0) \quad (\text{A.4})$$

and

$$(Q^*, T_g^*)_E (Q^*, T_2^*) \sim (Q^*, T_G^*)_E (Q^*, T_1). \quad (\text{A.5})$$

Eqs. A.4 and A.5 imply that T_0, T_1, T_2^* is a Q^* -standard sequence. Because $T_2^* < T_2$, a contradiction results with standard sequence invariance on \mathcal{T} . *QED*

Because $T_1 \sim T_1^*$, T_1 (and also T_1^*) is both the V and the V^* midpoint of T_0 and T_2 . Hence, on $\{T' \in \mathcal{T}: T^- < T' < T^+\}$, V and V^* midpoints are the same. With $V^* = \varphi \circ V$, the continuous function φ satisfies $\varphi((v_1 + v_2)/2) = (\varphi(v_1) + \varphi(v_2))/2$ on the interval $(V(T^-), V(T^+))$ around $V(T)$. It must be affine on this interval (Aczel, 1966 §2.1.3) and have second derivative 0 there, including at T .

The continuous and strictly increasing φ has second derivative 0 at all T in the interior of its domain $V(\mathcal{T})$. This implies that it is affine everywhere. Hence $V^*(T) = U(Q^*, T) = f(Q^*)V(T) + g(Q^*)$ for a positive $f(Q^*)$. This implies Eq. 4.2.

REMARK A.2. In this proof, we only used standard sequences in Eq. 4.1 with $Q^* = Q$. Hence the theorem remains valid if we define standard sequences only for $Q^* = Q$ in Eq. 4.1, and impose standard sequence invariance only for those standard sequences. The resulting condition is mathematically interesting because it is a common weakening of utility independence and standard sequence invariance, implying that the resulting modification of Theorem 4.2 is an immediate generalization of the theorems with utility independence in the literature. We chose the stronger version of standard sequence invariance in our main text because it is empirically more useful. \square

PROOF OF OBSERVATION 4.3. Substituting the null element in Eq. 4.2 shows that $g(Q)$ must be constant. It can be taken 0 because U is an interval scale. \square

PROOF OF OBSERVATION 5.2. Assume utility independence on a set of the form X_{\downarrow}^2 . This implies Eq. 3.3 for utility. This, in turn, implies utility independence on the whole domain of prospects because changing the deterministic level of some attributes amounts to an interval rescaling of utility, which does not affect preference. Utility independence on the whole domain trivially implies utility independence on the set X_{\downarrow}^2 . Hence Eq. 3.3 and the two versions of utility independence are equivalent.

Next assume standard sequence invariance on a set of the form X_{\downarrow}^2 . This implies Eq. 4.2 for utility. This, in turn, implies standard sequence invariance on every set isomorphic to a set X_{\downarrow}^2 . Hence Eq. 4.2 and the two versions of standard sequence invariance are equivalent.

REMARK A.3. Although we did not formally define standard sequences on larger domains, it can readily be seen that such versions are easy to obtain. Replacing the deterministic level of some attributes amounts to an interval rescaling of utility, which does not alter equal spacedness of utility on, for instance, comonotonic subsets under rank-dependent utility. \square

References

- Aczél, J. (1966). *Lectures on functional equations and their applications*. New York: Academic Press.
- Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science*, *46*, 1497-1512.
- Abdellaoui, M., Vossman, F., & Weber, M. (2005). Choice-based elicitation and decomposition of decision weights for gains and losses under uncertainty. *Management Science*, *51*, 1384-1399.
- Aczel, J. (1966). *Lectures on functional equations and their applications*. New York: Academic Press.
- Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica*, *21*, 503-546.
- Baron, J. (2008). *Thinking and deciding, 4th ed.* Cambridge: Cambridge University Press.
- Bier, V. M. & Connell, B. L. (1994). Ambiguity seeking in multi-attribute decisions: Effects of optimism and message framing. *Journal of Behavioral Decision Making*, *7*, 169-182.
- Bleichrodt, H., Cillo, A., & Diecidue, E. (2010). A quantitative measurement of regret theory. *Management Science*, *56*, 161-175.
- Bleichrodt, H. & Johannesson, M. (1997). The validity of QALYs: An empirical test of constant proportional tradeoff and utility independence. *Medical Decision Making*, *17*, 21-32.
- Bleichrodt, H. & Miyamoto, J. (2003). A characterization of quality-adjusted life-years under cumulative prospect theory. *Mathematics of Operations Research*, *28*, 181-193.

- Bleichrodt, H. & Pinto, J. L. (2000). A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management Science*, 46, 1485-1496.
- Bleichrodt, H. & Pinto, J. L. (2005). The validity of QALYs under non-expected utility. *Economic Journal*, 115, 533-550.
- Bleichrodt, H. & Quiggin, J. (1997). Characterizing QALYs under a general rank dependent utility model. *Journal of Risk and Uncertainty*, 15, 151-165.
- Bleichrodt, H., Schmidt, U., & Zank, H. (2009). Additive utility in prospect theory. *Management Science*, 55, 863-873.
- Bleichrodt, H., Wakker, P. P., & Johannesson, M. (1997). Characterizing QALYs by risk neutrality. *Journal of Risk and Uncertainty*, 15, 107-114.
- Booij, A. S. & van de Kuilen, G. (2009). A parameter-free analysis of the utility of money for the general population under prospect theory. *Journal of Economic Psychology*, 30, 651-666.
- Bouyssou, D. & Pirlot, M. (2003). Nontransitive decomposable conjoint measurement. *Journal of Mathematical Psychology*, 46, 677-703.
- Bouyssou, D. & Pirlot, M. (2004). A note on Wakker's cardinal coordinate independence. *Mathematical Social Sciences*, 48, 11-22.
- Casadesus-Masanell, R., Klibanoff, P., & Ozdenoren, E. (2000). Maxmin expected utility over Savage acts with a set of priors. *Journal of Economic Theory*, 92, 35-65.
- Chateauneuf, A. & Wakker, P. (1993). From local to global additive representation. *Journal of Mathematical Economics*, 22, 523-545.
- Conte, A., Hey, J. D., & Moffatt, P. G. (2011). Mixture models of choice under risk. *Journal of Econometrics*, 162, 79-88.

- Doctor, J. N., Bleichrodt, H., Miyamoto, J., Temkin, N. R., & Dikmen, S. (2004). A new and more robust test of QALYs. *Journal of Health Economics*, 23, 353-367.
- Doctor, J. N. & Miyamoto, J. (2003). Deriving quality-adjusted life-years (QALYs) from constant proportional time tradeoff and risk posture conditions. *Journal of Mathematical Psychology*, 47, 557-567.
- Dyckerhoff, R. (1994). Decomposition of multivariate utility functions in non-additive utility theory. *Journal of Multi-Criteria Decision Analysis*, 3, 41-58.
- Ebert, U. (2004). Social welfare, inequality, and poverty when needs differ. *Social Choice and Welfare*, 23, 415-448.
- Ellsberg, D. (1961). Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics*, 75, 643-669.
- Engel, Y. & Wellman, M. P. (2010). Multiattribute auctions based on generalized additive independence. *Journal of Artificial Intelligence Research*, 37, 479-525.
- Feeny, D. (2006). The multi-attribute approach to assessing health-related quality of life. In A. M. Jones (Eds.), *The Elgar companion to health economics* (pp. 359-370). Cheltenham, UK & Northampton MA, USA: Edward Elgar.
- Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., Depauw, S., Denton, M., & Boyle, M. (2002). Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Medical Care*, 40, 113-128.
- Fishburn, P. C. (1967). Methods of estimating additive utilities. *Management Science*, 13, 435-453.
- Fishburn, P. C. & Edwards, W. (1997). Discount-neutral utility models for denumerable time streams. *Theory and Decision*, 34, 139-166.

- Fishburn, P. C. & Rubinstein, A. (1982). Time preference. *International Economic Review*, 23, 677-694.
- Gajdos, T., Hayashi, T., Tallon, J.-M., & Vergnaud, J.-C. (2008). Attitude towards imprecise information. *Journal of Economic Theory*, 140, 27-65.
- Ghirardato, P. & Marinacci, M. (2001). Risk, ambiguity, and the separation of utility and beliefs. *Mathematics of Operations Research*, 26, 864-890.
- Gilboa, I. (1987). Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics*, 16, 65-88.
- Gilboa, I. & Schmeidler, D. (1989). Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics*, 18, 141-153.
- Gilboa, I., Schmeidler, D., & Wakker, P. P. (2002). Utility in case-based decision theory. *Journal of Economic Theory*, 105, 483-502.
- Guerrero, A. M. & Herrero, C. (2005). A semi-separable utility function for health profiles. *Journal of Health Economics*, 24, 33-54.
- Gul, F. (1991). A theory of disappointment aversion. *Econometrica*, 59, 667-686.
- Harvey, C. M. (1986). Value functions for infinite period planning. *Management Science*, 32, 1123-1139.
- Hurwicz, L. (1951). Some specification problems and applications to econometric models. *Econometrica*, 19, 343-344.
- Jaffray, J.-Y. (1994). Dynamic decision making with belief functions. In R. R. Yager, M. Fedrizzi and J. Kacprzyk (Eds.), *Advances in the Dempster-Shafer theory of evidence* (pp. 331-352). New York: Wiley.
- Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-291.

- Keeney, R. & Raiffa, H. (1976). *Decisions with multiple objectives*. New York: Wiley.
- Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, vol. 1*. New York: Academic Press.
- Loewenton, E. & Luce, R. D. (1966). Measuring equal increments of utility for money without measuring utility itself. *Psychonomic Science*, 6, 75-76.
- Loomes, G., Starmer, C., & Sugden, R. (2003). Do anomalies disappear in repeated markets? *Economic Journal*, 113, C153-C166.
- Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). Introduction to stated preference models and methods. In J. J. Louviere, D. A. Hensher and J. D. Swait (Eds.), *Stated choice methods: Analysis and applications* (pp. 20-33). Cambridge: Cambridge University Press.
- Luce, R. D. (2000). *Utility of gains and losses: Measurement-theoretical and experimental approaches*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Luce, R. D. & Fishburn, P. C. (1991). Rank- and sign-dependent linear utility models for finite first-order gambles. *Journal of Risk and Uncertainty*, 4, 29-59.
- Miyamoto, J. M. (1999). Quality-adjusted life-years (QALY) utility models under expected utility and rank dependent utility assumptions. *Journal of Mathematical Psychology*, 43, 201-237.
- Miyamoto, J. M. & Eraker, S. A. (1988). A multiplicative model of the utility of survival duration and health quality. *Journal of Experimental Psychology: General*, 117, 3-20.
- Miyamoto, J. M. & Wakker, P. P. (1996). Multiattribute utility theory without expected utility foundations. *Operations Research*, 44, 313-326.

- Miyamoto, J. M., Wakker, P. P., Bleichrodt, H., & Peters, H. J. M. (1998). The zero-condition: A simplifying assumption in QALY measurement and multiattribute utility. *Management Science*, 44, 839-849.
- Nau, R. F. (2006). Uncertainty aversion with second-order utilities and probabilities. *Management Science*, 52, 136-145.
- Pliskin, J. S., Shepard, D. S., & Weinstein, M. C. (1980). Utility functions for life years and health status. *Operations Research*, 28, 206-223.
- Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3, 323-343.
- Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
- Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, 57, 571-587.
- Schmidt, U. (2003). Reference dependence in cumulative prospect theory. *Journal of Mathematical Psychology*, 47, 122-131.
- Skiadas, C. (1997). Subjective probability under additive aggregation of conditional preferences. *Journal of Economic Theory*, 76, 242-271.
- Spencer, A. & Robinson, A. (2007). Test of utility independence when health varies over time. *Journal of Health Economics*, 26, 1003-1013.
- Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, 28, 332-382.
- Stigler, G. J. (1950). The development of utility theory. I. *Journal of Political Economy*, 58, 307-327.
- Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5, 297-323.

- Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, 95, 371-384.
- van de Kuilen, G. & Wakker, P. P. (2011). The midweight method to measure attitudes toward risk and ambiguity. *Management Science*, 57, 582-598.
- von Neumann, J. & Morgenstern, O. (1944). The theory of games and economic behavior. Princeton, NJ: *Princeton University Press*.
- von Winterfeldt, D. & Edwards, W. (1986). Decision analysis and behavioral research. Cambridge: *Cambridge University Press*.
- Wakker, P. P. (1984). Cardinal coordinate independence for expected utility. *Journal of Mathematical Psychology*, 28, 110-117.
- Wakker, P. P. (1991). Additive representations on rank-ordered sets. I. The algebraic approach. *Journal of Mathematical Psychology*, 35, 501-531.
- Wakker, P. P. (2010). Prospect theory: For risk and ambiguity. Cambridge UK: *Cambridge University Press*.
- Wakker, P. P. & Deneffe, D. (1996). Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science*, 42, 1131-1150.
- Wakker, P. P. & Tversky, A. (1993). An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty*, 7, 147-176.
- Wald, A. (1950). *Statistical decision functions*. New York: Wiley.

New Tests of QALYs when Health Varies over Time

Summary

This chapter performs new tests of the QALY model when health varies over time. Our tests do not involve confounding assumptions and are robust to violations of expected utility. The results support the use of QALYs at the aggregate level, i.e. in economic evaluations of health care. At the individual level, there is less support for QALYs. The individual data are, however, largely consistent with a more general QALY-type model that remains tractable for applications.

Keywords: QALYs, economic evaluation of health care, decision under risk, nonexpected utility.

⁹ This chapter was published as Bleichord, H., Filko, M. (2008). New Tests of QALYs when Health Varies over Time. *Journal of Health Economics* 27, 1237 – 1249. We are grateful to Anirban Basu, Willard Manning, Elly Stolk, Peter Wakker, and three anonymous referees for helpful comments on a previous version of this paper. Han Bleichrodt's research was supported by a research grant from the Netherlands Organization for Scientific Research (NWO). Martin Filko's research was supported by a grant from DSW.

3.1 Introduction

Quality-adjusted life-years (QALYs) are the most widely used measure of health in economic evaluations of health care. According to the QALY model, the utility of a health profile equals the sum of the utilities of its constituent health states. The popularity of QALYs can be explained by their tractability and intuitive appeal: QALYs are easy to use and easy to explain to policy makers. A drawback of QALYs may be that they are too simple and do not represent people's preferences for health in a reliable manner. An obvious danger of using an unreliable measure in economic evaluations of health care is that treatment recommendations and reimbursement decisions are made that do not represent people's interests.

Several studies have tested the validity of QALYs when health states are chronic (for an overview see Bleichrodt and Pinto-Prades, 2006). Less evidence exists on the validity of QALYs for the more realistic case where health varies over time. Most of the existing studies tested the validity of QALYs by comparing the directly elicited utility of a health profile with the indirect utility that is obtained by adding the utilities of the independently rated constituent health states. The evidence from these studies is mixed with some studies finding large and significant differences (e.g. Richardson, Hall, and Salkeld, 1996) and others finding only small and typically insignificant differences (Mackeigan, O'Brien, and Oh, 1999, Brazier et al., 2006). The performance of the QALY model is better at the aggregate level than at the individual level (e.g. Kuppermann et al., 1997) although Krabbe and Bonsel (1998) found that only a small proportion of their subjects violated additivity.

There are several problems with the above method for testing the validity of QALYs. A first problem is that confounding assumptions must be made, in particular about the discounting of future health. All the above studies assumed constant discounting. Empirical evidence abounds, however, that the descriptive record of constant discounting is poor and that people deviate from it systematically (Frederick, Loewenstein, and O'Donoghue, 2002, van der Pol and Cairns, 2002). The problem of confounding assumptions is that when a difference between the direct and the indirect

valuation of a profile is observed we do not know what is causing this difference and, hence, no information is obtained on how QALYs might be improved.

A second problem is that the valuation of health profiles and the valuation of health states involve different experimental stimuli and may, therefore, invoke different cognitive processes. Consequently, they may be susceptible to different decision biases. In particular, a cognitively demanding task such as the valuation of health profiles may induce the use of simplifying heuristics.

A third problem arises if the measured health utilities are biased. Many of the abovementioned studies used the standard gamble. It is well known that the standard gamble leads to utilities that are too high (van Osch et al., 2004, Bleichrodt et al., 2007, Doctor, Bleichrodt, and Lin, forthcoming). This upward bias is only present once in the direct valuation of the health profiles, but affects the valuation of each of the constituent health states and, hence, it is present more than once in the indirect valuation of the health profiles based on the utilities of their constituent health states. Consequently, we would expect that the estimation of the utility of a health profile from its constituent health states exceeds the direct valuation of the profile when the standard gamble is used and this is indeed what is typically observed.

The above problems can be avoided by testing the preference conditions on which QALYs are based. This approach was adopted by Treadwell (1998), who tested preference independence, and Spencer and Robinson (2007), who tested utility independence. Preference independence and utility independence are implied by the QALY model, i.e. they are necessary conditions for the QALY model. Both Treadwell (1998) and Spencer and Robinson (2007) found that the condition they tested was generally supported. The support for these conditions does not imply, however, that the QALY model holds as the conditions are also consistent with other, more general, decision models. To obtain conclusive evidence on the validity of QALYs, conditions must be tested that are both implied by the QALY model and that imply the QALY model, i.e. conditions that are both necessary and sufficient.

Spencer (2003) tested such a condition. She observed a violation of this condition at the individual level, but the violation was not systematic and might just be due to noise. Spencer's test is only valid if people behave according to expected utility. It is

well known, however, that people systematically deviate from expected utility (Starmer, 2000). Hence, it cannot be excluded that the violations of the QALY model that Spencer observed reflected violations of expected utility rather than violations of the QALY model. To wit, while many studies observed violations of the QALY model for chronic health states under expected utility, Doctor et al. (2004) found no violations of the QALY model when violations of expected utility were taken into account.

In this chapter we provide new tests of the QALY model when health varies over time. Like Treadwell, (1998), Spencer (2003), and Spencer and Robinson (2007) we test preference conditions and, hence, our tests are not affected by the problems surrounding the comparison between direct and indirect valuations of health profiles. We test two conditions. The first condition, generalized marginality, is the central condition underlying the QALY model and implies that health profiles can be evaluated additively. Hence, like Spencer (2003) our test is both necessary and sufficient for the QALY model. An important difference with Spencer (2003) is that our test does not assume expected utility but is valid under a more general utility model that includes many of the theories of decision under risk that exist today. Hence, our tests are robust to violations of expected utility.

As generalized marginality is a restrictive condition and we could well imagine people violating it, we also tested utility independence. Utility independence is less restrictive than generalized marginality and it does not imply the QALY model. As will be explained in Section 2, utility independence still implies a model that is tractable and that can be used in practical applications. Spencer and Robinson (2007) also tested utility independence. Our experimental protocol differed in several respects from the protocol used by Spencer and Robinson (2007) and, hence, our data on utility independence complement Spencer and Robinson's analysis. Taken together the two studies provide insight into the validity of utility independence, an important condition for preference modeling and utility measurement.

The chapter is structured as follows. Section 2 provides theoretical background and explains generalized marginality and utility independence. Section 3 describes the

design of an experiment that aimed to test these conditions and Section 4 its results. Section 5 discusses the results and concludes.

3.2 Background

Let $q = (q_1, \dots, q_T)$ denote a *health profile* where q_t stands for the health state at period t and T denotes the number of periods of survival. We assume that all health states are better than death. In our experiment we will only consider health profiles consisting of three periods and, hence, we will take $T=3$ in what follows. A *prospect* $(p;q;r)$ gives health profile q with probability p and health profile r with probability $1 - p$.¹⁰ Throughout the chapter we will only use prospects involving at most two different health profiles.

A preference relation \succsim is given over the set of prospects. The conventional notation $>$ and \sim is used to denote strict preference and indifference. By restricting attention to *constant prospects*, i.e. prospects for which $q = r$ or for which $p = 0$ or $p = 1$, a preference relation over health profiles can be defined, which we also denote by \succsim . It is implicit in the notation $(p;q;r)$ that health profile q is at least as good as health profile r ($q \succsim r$), i.e. all prospects are *rank-ordered*.

We assume that a prospect $(p;q;r)$ can be evaluated through

$$\pi U(q) + (1 - \pi)U(r) \tag{1}$$

and choices and preferences correspond with this evaluation. In Eq.1, π is the decision weight assigned to the health profile q that obtains with probability p . This decision weight is entirely general. It depends on the probability p but we assume nothing about the way in which it depends on p . We will refer to Eq.1 as *general rank-dependent utility* (GRU). Equation 1 is consistent with many theories of decision under risk. For example, if $\pi = p$ then Eq.1 reduces to expected utility. If $\pi = w(p)$,

¹⁰ This means that there is an event E with probability p such that x obtains under E and y obtains under the complement of E . That is, we assume richness of the set of events.

with w a probability weighting function¹¹ then Eq.1 reduces to rank-dependent utility (Quiggin, 1981). If $\pi = 0$ then all weight is given to the worst health profile and Eq.1 corresponds to maximin. Eq. 1 was first suggested by Miyamoto (1988) and was subsequently used by Miyamoto and Wakker (1996) and Bleichrodt and Quiggin (1997).

Under the *QALY model* the function $U(\cdot)$ in Eq.1 is additive:

$$U(q) = \sum_{t=1}^T V_t(q_t) \quad (2)$$

where the functions V_t can be period-specific. Often a more restrictive QALY model is used where the functions V_t are common for all periods and a constant discount factor is applied to all periods:

$$U(q) = \sum_{t=1}^T \delta^{t-1} V(q_t). \quad (3)$$

The focus in this chapter is on Eq.2, which captures the essential idea of QALYs of additivity over time. Bleichrodt and Gafni (1996) showed how Eq.3 can be obtained from Eq.2 by adding one preference condition.

Let $a_i v_j q$ denote the health profile q with health state q_i replaced by a_i and health state q_j replaced by v_j , $i, j \in \{1, 2, 3\}$, $i \neq j$. For example, if $i = 1$, $j = 2$, then $a_i v_j q = (a_1, v_2, q_3)$. Consider the following condition:

Definition 1: the preference relation \succsim satisfies *generalized marginality* when for all $i, j \in \{1, 2, 3\}$, $i \neq j$, and for all health profiles q , health states a, b, c, d, v, w, x, y , and for all p :

$$(p: a_i v_j q; b_i w_j q) \sim (p: c_i v_j q; d_i w_j q) \Leftrightarrow (p: a_i x_j q; b_i y_j q) \sim (p: c_i x_j q; d_i y_j q).$$

Consider first the indifference $(p: a_i v_j q; b_i w_j q) \sim (p: c_i v_j q; d_i w_j q)$. In terms of marginal probabilities the two prospects are almost identical except that the first one gives a

¹¹ That is, w is increasing (if $p > q$ then $w(p) > w(q)$) and satisfies $w(0)=0$ and $w(1)=1$.

probability p of health state a in period i and a probability $(1-p)$ of health state b in period i and the second a probability p of health state c in period i and a probability $1-p$ of health state d in period i . Both prospects give a probability p of health state v in period j and a probability $1-p$ of health state w in period j and a probability 1 of getting q in the remaining period k . Hence, in terms of marginal probabilities the indifference implies that getting a_i with probability p and b_i with probability $1-p$ is just sufficient to offset getting c_i with probability p and d_i with probability $1-p$.

The only difference in the second indifference, $(p:a_i x_j q; b_i y_j q) \sim (p:c_i x_j q; d_i y_j q)$, is that there is a change in what happens in time period j : health state v is replaced by health state x and health state w by health state y . The latter change is such that the two prospects still yield the same marginal probability distribution over what happens in time period j : in both prospects there is a probability p of obtaining health state x in period j and a probability $1-p$ of obtaining health state y . Generalized marginality says that this change should not affect indifference. Getting a_i with probability p and b_i with probability $1-p$ should still be just sufficient to offset getting c_i with probability p and d_i with probability $1-p$. Essentially, generalized marginality says that preferences depend only on marginal probabilities (hence the term marginal in generalized marginality) and not on the joint probability distribution.

An example may clarify the restrictiveness of generalized marginality. Let there be four health states: good health, fair health, poor health, and very poor health. Suppose that a decision maker is indifferent between:

$(\frac{1}{2}: (\text{good, good, poor}); (\text{poor, fair, poor}))$ and $(\frac{1}{2}: (\text{fair, good, poor}); (\text{fair, fair, poor}))$.

In both prospects there is a possibility of good health. In the first one the probability of good health is higher but there is also a higher probability of poor health. Generalized marginality then implies that the decision maker should also be indifferent between:

$(\frac{1}{2}: (\text{good, poor, poor}); (\text{poor, very poor, poor}))$ and $(\frac{1}{2}: (\text{fair, poor, poor}); (\text{fair, very poor, poor}))$.

It is, however, conceivable that a decision maker is not indifferent between these latter two prospects. For example, he may prefer the first prospect because this gives at least some time in good health, whereas in the second prospect both health profiles are pretty bad. The example shows that there are no a priori reasons why people should or would behave according to generalized marginality.

It is easy to show that under the GRU model, the QALY model (Eq.2) implies generalized marginality. To improve the understanding of what generalized marginality entails, we give the proof in the main text. Let $k \neq i, j$. Under GRU and the QALY model, $(p:a_i v_j q; b_i w_j q) \sim (p:c_i w_j q; d_i x_j q)$ implies that

$$\begin{aligned} & \pi(V_i(a_i) + V_j(v_j) + V_k(q_k)) + (1-\pi)(V_i(b_i) + V_j(w_j) + V_k(q_k)) = \\ & \pi(V_i(c_i) + V_j(v_j) + V_k(q_k)) + (1-\pi)(V_i(d_i) + V_j(w_j) + V_k(q_k)) \end{aligned} \quad (4a)$$

or

$$\pi V_i(a_i) + (1-\pi)V_i(b_i) = \pi V_i(c_i) + (1-\pi)V_i(d_i) \quad (4b)$$

Eq. 4b implies that

$$\begin{aligned} & \pi(V_i(a_i) + V_j(x_j) + V_k(q_k)) + (1-\pi)(V_i(b_i) + V_j(y_j) + V_k(q_k)) = \\ & \pi(V_i(c_i) + V_j(x_j) + V_k(q_k)) + (1-\pi)(V_i(d_i) + V_j(y_j) + V_k(q_k)). \end{aligned}$$

and substitution of $\pi V_i(a_i) + (1-\pi)V_i(b_i) = \pi V_i(c_i) + (1-\pi)V_i(d_i)$ implies $(p:a_i x_j q; b_i y_j q) \sim (p:c_i x_j q; d_i y_j q)$.

Bleichrodt and Quiggin (1997, Theorem 4) showed that under GRU, the QALY model not only implies generalized marginality, but generalized marginality also

implies the QALY model.¹² Hence, generalized marginality is the central condition of the QALY model.

We next define utility independence. Let J be a subset of $\{1,2,3\}$ and let q and k be two health profiles. By k_Jq we denote the health profile q with health state q_j replaced by health state k_j for all j in J . For example, if $J = \{1,3\}$ then $k_Jq = (k_1, q_2, k_3)$.

Definition 2: The preference relation \succsim satisfies *utility independence* if for all subsets J of $\{1,2,3\}$, for all health profiles k,l,m,n,q,r , and for all probabilities p :

$$(p:k_Jq; l_Jq) \sim (p:m_Jq; n_Jq) \Leftrightarrow (p:k_Jr; l_Jr) \sim (p:m_Jr; n_Jr).$$

That is, if all health profiles in the prospects under comparison have common health states outside J preferences do not depend on what these common health states are.

Consider, again, the example given before. If the decision maker is indifferent between

$(\frac{1}{2}: (\text{good, good, poor}); (\text{poor, fair, poor}))$ and $(\frac{1}{2}: (\text{fair, good, poor}); (\text{fair, fair, poor}))$

then utility independence says that he should also be indifferent between

$(\frac{1}{2}: (\text{good, good, good}); (\text{poor, fair, good}))$ and $(\frac{1}{2}: (\text{fair, good, good}); (\text{fair, fair, good}))$,

where we changed the common outcome in the third period from poor to good. It is, however, conceivable that the decision maker is not indifferent between the second pair of prospects. He may now, for instance, prefer the second prospect because it does not carry the risk of spending time in poor health. Like generalized marginality, utility independence is not a priori obviously fulfilled.

¹² This result requires richness of the set of health states, i.e. indifferences can be obtained by variations in the health states. For empirical testing it is easier to vary probabilities and, therefore, we assumed a rich set of events (see footnote 1). Whether the presence of a rich set of events implies that the result of Bleichrodt and Quiggin (1997) still hold without richness of the set of health states assumed is an open question.

Utility independence is less restrictive than generalized marginality: generalized marginality implies utility independence but the reverse is not true. Miyamoto and Wakker (1996, Theorem 4) showed that if utility independence holds but generalized marginality is violated then

$$U(q) = \prod_{t=1}^T V_t(q_t), \quad (5)$$

i.e., utility is multiplicative. Equation 5 is still tractable. Consequently, not all is lost when generalized marginality is violated and, in the face of possible violations of generalized marginality, it is important to test utility independence.

Guerrero and Herrero (2005) further relaxed utility independence and only imposed it for initial health states. They showed that even then a reasonably tractable model results. Their condition is hard to test empirically because it involves dynamic decisions and tests of their model require the comparison of choices made at different points in time. We do not consider their model in this chapter.

3.3 Experiment

Test—The aim of the experiment was to test generalized marginality and utility independence. The general structure of our tests of generalized marginality was as follows. First we elicited the probability p_{vw} such that a subject was indifferent between prospects of the type $(p_{vw}:a_i v_j q; b_i w_j q)$ and $(p_{vw}:c_i v_j q; d_i w_j q)$. Then we elicited the probability p_{xy} such that subjects were indifferent between $(p_{xy}:a_i x_j q; b_i y_j q)$ and $(p_{xy}:c_i x_j q; d_i y_j q)$ with x and y different from v and w . Under generalized marginality we should observe that $p_{vw} = p_{xy}$ except for random error.

To test for utility independence we first elicited the probability p_q such that subjects were indifferent between $(p_q:k_j q; l_j q)$ and $(p_q:m_j q; n_j q)$ for a given subset J . Then we elicited the probability p_r such that subjects were indifferent between $(p_r:k_j r; l_j r)$ and $(p_r:m_j r; n_j r)$ with r different from q . Under utility independence we should observe that $p_q = p_r$ except for random error.

Subjects—Subjects were 60 students (30 female, median age of all subjects 22 years) from Erasmus University. They were paid a flat fee of €10. Prior to the actual experiment, the experimental design was tested and fine-tuned in several pilot sessions.

Procedure—The experiment was run on a computer in personal interview sessions. To reduce errors and to ensure that subjects could focus entirely on the experimental questions, all responses were entered into the computer by the interviewer. Subjects were told that there were no right or wrong answers and that we were only interested in their preferences. Experimental sessions lasted 40 minutes on average and consisted of three parts: instructions and practice questions, data collection for the experiment reported here, and data collection for an unrelated experiment. Subjects took approximately 15-20 minutes to answer the questions for this experiment.

All indifference probabilities were elicited through a series of choices. Each choice question corresponded to an iteration in a bisection process, which is described in Appendix B. A choice-based elicitation procedure was used because previous studies observed that inferring indifferences from a series of choices leads to fewer inconsistencies than asking subjects directly for their indifference value (see Luce (2000) for a review). The iterative process ended when the absolute difference in probability between successive steps in the iteration was less than 5 percentage points. We learned from the pilot sessions that it was unrealistic to determine the probabilities with more precision. At the end of each iteration process we repeated the first question of the iteration process. If subjects gave the same answer to this repeated choice question then we moved on to the next elicitation. If not, the iteration process for this elicitation was started anew. The aim of repeating the first choice in the iteration process was to reduce the impact of decision errors.

Stimuli— Subjects were asked to make a choice between two prospects consisting of two health profiles, neutrally labeled A and B. Figure 1 shows the way the prospects were displayed on the computer screen.

Figure 1: Example of an experimental question

<p>Option A</p>	<p>Probability of</p> <p style="text-align: center;">74 %</p>	<p>Probability of</p> <p style="text-align: center;">26 %</p>											
	<p>Living for</p> <table border="1" style="width: 100%; text-align: center;"> <tr> <td>20 years</td> <td>20 years</td> <td>20 years</td> </tr> <tr> <td style="background-color: #f4cccc;">M</td> <td style="background-color: #f4cccc;">M</td> <td style="background-color: #f4cccc;">N</td> </tr> </table>	20 years	20 years	20 years	M	M	N	<p>Living for</p> <table border="1" style="width: 100%; text-align: center;"> <tr> <td>20 years</td> <td>20 years</td> <td>20 years</td> </tr> <tr> <td style="background-color: #f4cccc;">M</td> <td style="background-color: #f4cccc;">N</td> <td style="background-color: #f4cccc;">N</td> </tr> </table>	20 years	20 years	20 years	M	N
20 years	20 years	20 years											
M	M	N											
20 years	20 years	20 years											
M	N	N											
<p>Option B</p>	<p>Probability of</p> <p style="text-align: center;">74 %</p>	<p>Probability of</p> <p style="text-align: center;">26 %</p>											
	<p>Living for</p> <table border="1" style="width: 100%; text-align: center;"> <tr> <td>20 years</td> <td>20 years</td> <td>20 years</td> </tr> <tr> <td style="background-color: #d9ead3;">K</td> <td style="background-color: #f4cccc;">M</td> <td style="background-color: #f4cccc;">N</td> </tr> </table>	20 years	20 years	20 years	K	M	N	<p>Living for</p> <table border="1" style="width: 100%; text-align: center;"> <tr> <td>20 years</td> <td>20 years</td> <td>20 years</td> </tr> <tr> <td style="background-color: #f4cccc;">N</td> <td style="background-color: #f4cccc;">N</td> <td style="background-color: #f4cccc;">N</td> </tr> </table>	20 years	20 years	20 years	N	N
20 years	20 years	20 years											
K	M	N											
20 years	20 years	20 years											
N	N	N											

Health profiles consisted of three periods of 20 years each. Hence, the total length of the profiles was 60 years, which corresponded to the life-expectancy of our subjects. We used only three periods to keep the tasks as simple as possible. We used periods of 20 years because these more or less correspond to different life stages. The health states constituting the health profiles were selected from a set of four EuroQol health states. We selected only moderate health states because these can be imagined more easily by a healthy population like our subjects. Another reason to use moderate health states was to avoid considerations of maximal endurable time (Stalmeier, Wakker, and Bezembinder, 1997). Health states were labeled using capital letters from the middle of the alphabet, minimizing potential distorting associations (for example using the letter D might be associated with the outcome death). The ordering of the health states was obvious in the sense that more preferred health states scored at least as good on each EuroQol dimension as less preferred health states and strictly better on at least one dimension. The ordering of the health states corresponded with the alphabetical order.

Table 1: Description of health states used in the experiment

Label	Color	EQ code	EQ utility
K	Green	11111	1.000
L	Yellow	11121	0.850
M	Orange	11122	0.722
N	Red	12222	0.551

Health states were printed on separate cards and were assigned colors in an intuitive order (green the best, red the worst, etc.). The use of color-coding aimed to facilitate decision making by reminding the subjects of the relative attractiveness of the health states. The EuroQoL system was introduced in the initial instructions and, throughout the experiment, subjects had the cards describing the health states in front of them. Health states are summarized in Table 1 and the cards that were handed to the subjects are reproduced in Appendix A. The final column of Table 1 displays the utility of the health states according to the EuroQol algorithm (Dolan, 1997).

It is crucial for our tests that the prospects are rank-ordered. To ensure this we selected the prospects such that one profile yielded in each period a health state that was always at least as good as the other profile. To help subjects understand the ranking of health profiles in each of the choices they faced, we asked them - before the bisection procedure for a particular question started - to rank the four health profiles involved in the question from the best to the worst with ties allowed. No violations of rank-ordering were observed.

We performed three tests of generalized marginality and four tests of utility independence. The tests of generalized marginality are displayed in Table 2, those of utility independence in Table 3. All tests consisted of two parts. The first part of a test is indicated with the Roman number I in the tables, the second part with the number II. The prospect mentioned first was displayed to the subjects as option A, the other as option B. KMN denotes a profile that gives health state K for the first 20 years, health state M for the next 20 years and health state N for the final 20 years. Outcomes that were varied between the two parts of each test are underlined. Note that in the fourth test of utility independence two common outcomes were changed.

Table 2: Tests of generalized marginality

Test	Part	Question
1	I	(p:MM <u>N</u> ; M <u>NN</u>) vs. (p:K <u>MN</u> ; N <u>NN</u>)
	II	(p:M <u>KN</u> ; M <u>MN</u>) vs. (p:K <u>KN</u> ; N <u>MN</u>)
2	I	(p:K <u>LM</u> ; K <u>MN</u>) vs. (p:K <u>KM</u> ; K <u>NN</u>)
	II	(p:K <u>LL</u> ; K <u>MM</u>) vs. (p:K <u>KL</u> ; K <u>NM</u>)
3	I	(p:K <u>LM</u> ; <u>LLM</u>) vs. (p:K <u>LL</u> ; <u>LLN</u>)
	II	(p:K <u>LM</u> ; <u>MLL</u>) vs. (p:K <u>LL</u> ; <u>MLN</u>)

The order in which the questions were asked was arbitrary with the restriction that the two parts of a given test were never offered consecutively. Interspersing trials were implemented to prevent subjects from forming a match that would guide answers.

The experiment ended with two consistency tests in which subjects repeated the first part of the first test of generalized marginality (GM1-I) and the second part of the third test of utility independence (UI3-II).

Table 3: Tests of utility independence

Test	Part	Question
1	I	(p:L <u>LM</u> ; M <u>NM</u>) vs. (p:L <u>MM</u> ; M <u>MM</u>)
	II	(p:L <u>LN</u> ; M <u>NN</u>) vs. (p:L <u>MN</u> ; M <u>MN</u>)
2	I	(p: <u>KKL</u> ; <u>KNN</u>) vs. (p: <u>KML</u> ; <u>KMN</u>)
	II	(p: <u>LKL</u> ; <u>LNN</u>) vs. (p: <u>LML</u> ; <u>LMN</u>)
3	I	(p: <u>LKN</u> ; <u>LNN</u>) vs. (p: <u>LMN</u> ; <u>LMN</u>)
	II	(p: <u>NKN</u> ; <u>NNN</u>) vs. (p: <u>NMN</u> ; <u>NMN</u>)
4	I	(p: <u>KML</u> ; <u>KMN</u>) vs. (p: <u>KMM</u> ; <u>KMM</u>)
	II	(p: <u>MLL</u> ; <u>MLN</u>) vs. (p: <u>MLM</u> ; <u>MLM</u>)

Spencer and Robinson (2007) also tested utility independence and found support for it in 6 out of 8 tests. The main difference between their study and ours is that they asked directly for indifferences whereas we used a choice-based elicitation method.¹³ It is well known from the literature that matching and choice invoke different cognitive processes (Tversky, Sattath, and Slovic, 1988). If the two studies were to give similar results in spite of the different response modes used then this would offer convincing evidence in favor of utility independence.

¹³ Other differences are that they used group sessions of 10-20 subjects whereas we used personal interviews, they used pen and paper whereas our experiment was computer-run, they used five periods of five years whereas we used three periods of 20 years, and they used health states ranging from normal health to death whereas we used only moderate health states. Finally, they asked the two parts of the tests of utility independence consecutively in their first experiment, but not in their second experiment, which randomized the three tests that were used.

Analysis—We used both parametric (t-test) and nonparametric (Wilcoxon) tests to test for significance of differences. Unless a difference was observed we only report the parametric results. Because we performed many different tests, there is a danger of finding significant differences just by chance. To reduce this danger we used a significance level of 1% in the statistical tests reported below. Using 5% instead did not affect our conclusions much.

Our sample size with a standard deviation of 0.15 would have enough power to detect a difference in aggregate values of 0.08 ($\alpha = 0.01$, $1 - \beta = 0.90$). In the literature a difference of 0.10 is often considered meaningful and important in decision-making contexts (O'Brien and Drummond, 1994). Hence, the power of our study was satisfactory.

3.4 Results

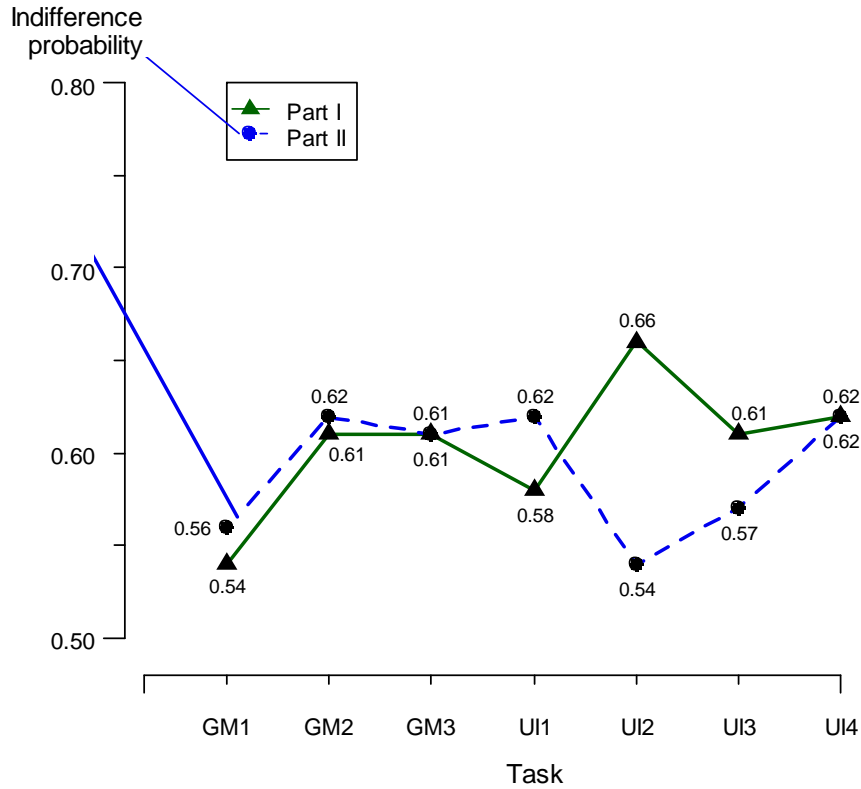
Consistency—Three subjects were excluded either for not cooperating or for targeting towards 0% and 100% in each question. This left 57 subjects in the final analysis. The consistency tests yielded mixed results. In the test for GM1-I the median probabilities were 0.54 in the original test and 0.58 in the retest. The difference was not significant ($p = 0.064$). The median of the individual absolute differences between test and retest was 0.06. In the test for UI3-II the median probabilities were 0.62 in the original experiment and 0.60 in the retest. In this case, the difference was significant, however (t-test $p = 0.005$ and Wilcoxon test $p = 0.011$). The median of the individual absolute differences between test and retest was 0.04. The Pearson correlation coefficients between test and retest were in both cases 0.63.

Few subjects had to restart the iteration process because they reversed their first choice. For the median subject this happened in just 1 out of 16 tests. In total, the proportion of reversals was 11.3%. This suggests that errors were rare when probabilities differed substantially from their indifference values, which was usually the case in the initial choices. By comparison, reversals up to 33% are common in choice experiments (Stott, 2006).

Aggregate results—Figure 2 displays the medians of the elicited indifference probabilities. The means were similar. The figure shows support for generalized marginality: for all three tests the median probabilities for both parts of the test were very close. None of the differences was significant ($p > 0.60$ in all three tests). The correlation coefficients between the two parts were, however, rather low. Correlation was only fair for GM1 (0.27) and GM2 (0.30) and was moderate (0.52) for GM3.

Figure 2 shows that the differences between the median probabilities were generally larger in the tests of utility independence than in the tests of generalized marginality. There appears to be no systematic pattern in the medians, however. In the first test of utility independence (UI1), the indifference probability was larger in the second part, in UI2 and UI3 it was larger in the first part and in UI4 there was no difference. We could not reject utility independence in three out of four tests (UI1, UI3, and UI4). The only exception is the second test. Here the difference in elicited probabilities is significant (t-test $p = 0.008$ and Wilcoxon test $p = 0.016$). Correlations between the two parts of the tests are higher than for generalized marginality and are moderate in all tests (0.44 in UI1, 0.46 in UI2, 0.51 in UI3, and 0.55 in UI4).

Figure 2: Median probabilities in the two parts of the tests



Individual results— Figure 3 shows the means and the medians of the individual absolute differences between the elicited probabilities in the two parts of each test. It should be kept in mind when interpreting these results that our choice-based procedure was terminated when the absolute difference in probabilities between successive iterations was less than 0.05 and the indifference value was set equal to the midpoint of the elicited interval (see Appendix B for details). This implies that there could be a maximum difference of 0.04 between the elicited probability and the true probability. Hence, when we compare the probabilities between the two parts of a test a difference of 0.08 might in theory have been caused by our elicitation procedure.

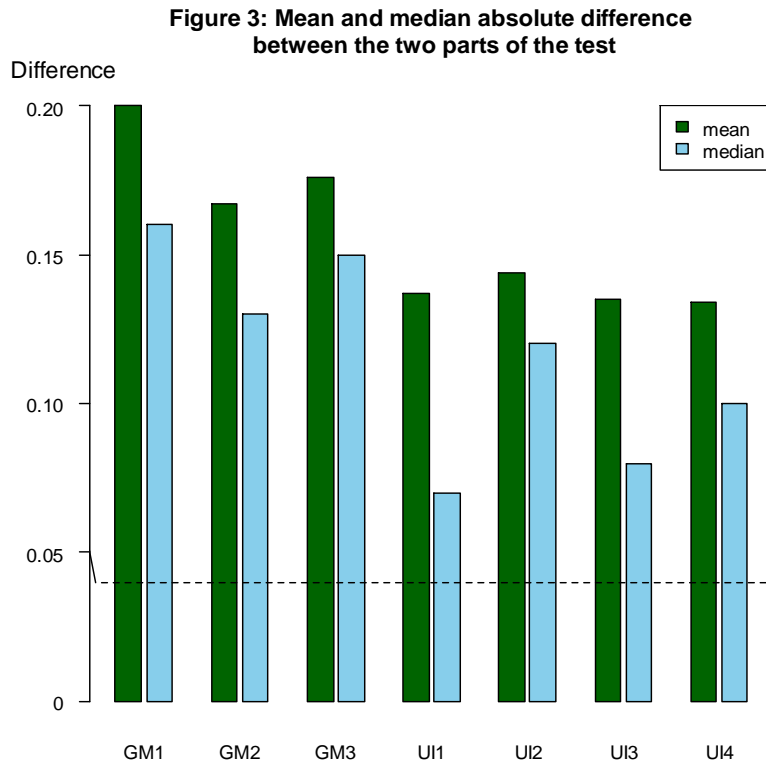


Table 4 shows the probabilities of observing a given absolute difference under our elicitation method when in reality no difference exists. The table was constructed under the assumption that any value from the elicited indifference interval was as likely to be the true indifference value. Table 4 displays that observing a difference of 0.08 due to our elicitation method alone was very unlikely: the chance of it happening was less than 0.001. The chance that our elicitation method led to an observed difference larger than 0.04 was, in fact, only 0.077. Differences up to 0.04 were however plausible. To illustrate this, we have plotted the value of 0.04 by a dotted line in Figure 3. Up to this line differences might reasonably be attributed to our elicitation method. Above it they cannot be explained by our elicitation method alone. The figure shows that both for utility independence and, in particular, for generalized marginality the observed absolute differences are clearly larger than 0.04 and, hence, they are not just products of imprecision in our elicitation method.

Table 4: Probability of a difference in probability being caused by elicitation method

Difference	Probability
0	0.148
0.01	0.276
0.02	0.226
0.03	0.166
0.04	0.106
0.05	0.053
0.06	0.019
0.07	0.004
0.08	<0.001

Another thing to take into account when considering Figure 3 is that subjects' preferences are likely to be imprecise (Dubourg, Jones-Lee, and Loomes, 1994, Butler and Loomes, 2007). The choices we asked our subjects to make were not easy and subjects had to compare probabilities, health states, and the timing and sequence of the health states simultaneously. When faced with complex choices it seems unrealistic to expect that subjects always have clear preferences between the two options. If subjects were, for instance, only able to distinguish between probabilities when they differed by at least 0.05 then an observed difference of 0.14 between the two parts of the tests could in theory be entirely caused by our elicitation method and preference imprecision. This value was however extremely unlikely: it had a probability of 0.0002. Of course we do not know exactly how much of the differences

that we observed were actually caused by preference imprecision but it is likely to have played at least some role in the observed differences. The medians of the individual differences between the test and retest for GM1-I and UI3-II may give some indication of this imprecision. They were 0.06 and 0.04 respectively and the value of 0.05 that we used in the above example was selected because it is the midpoint of these two values. It was also the median imprecision that was observed by Bleichrodt and Johannesson (1997) who made an attempt to quantify preference imprecision in health utility measurement.

Table 5: Classification of subjects based on the number of times they violated generalized marginality using different thresholds

Number of violations	Threshold	
	0.08	0.13
0	4	12
1	12	13
2	22	21
3	19	11

Table 5 presents a classification of our subjects based on the number of times that their responses exceeded a given threshold in the tests of generalized marginality. To account for the possibility of differences due to the elicitation procedure and due to preference imprecision we report the results for two different thresholds: 0.08 and 0.13. The table shows, for example, that there were only 4 subjects for whom the difference between the two parts of the tests was less than 0.08 in all three tests of generalized marginality. The conclusions drawn from the table depend on the threshold that is deemed plausible. Regardless of the threshold used, it seems safe to

conclude that a substantial proportion of our subjects violated generalized marginality and, consequently, the QALY model.

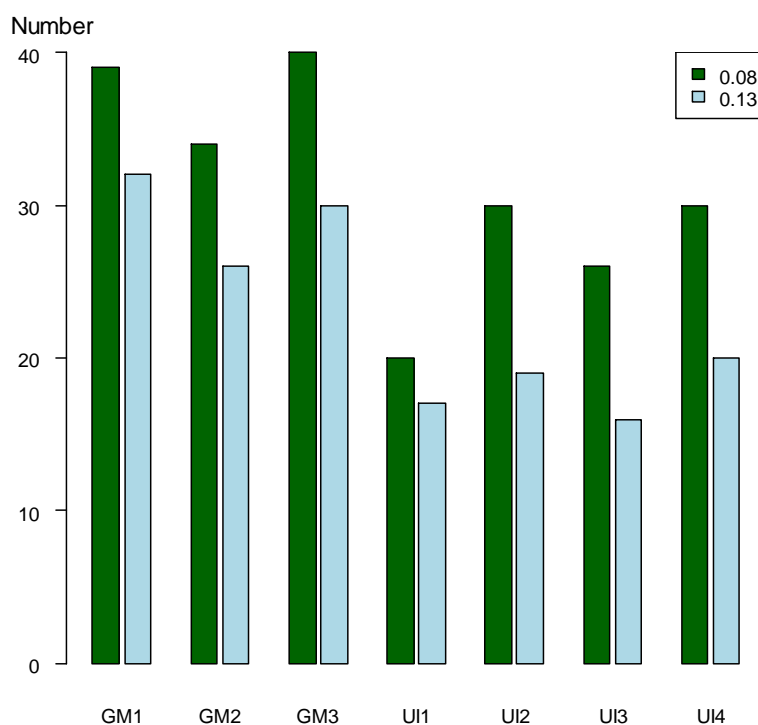
Table 6: Classification of subjects based on the number of times they violated utility independence using different thresholds

Number of violations	Threshold	
	0.08	0.13
0	4	14
1	16	21
2	23	15
3	12	7
4	2	0

Table 6 presents the same classification for the tests of utility independence. A comparison between Tables 5 and 6 reveals that there is more support for utility independence than for generalized marginality at the individual level. If we use a threshold of 0.13, over 60% of the subjects satisfied utility independence in at least three out of four tests. There were hardly any subjects who violated utility independence in each test, whereas the proportion of subjects violating generalized marginality in each test was substantial.

Figure 4 shows the number of subjects violating a particular test. The figure confirms that violations were more common in the tests of generalized marginality than in the tests of utility independence. The figure also shows that violations were not confined to one particular test but occurred in all tests.

Figure 4: Number of subjects violating a test for two different thresholds



3.5 Discussion

Main findings—We have performed new tests of the QALY model and a generalization thereof. Our tests do not require additional confounding assumptions, for example about discounting, and take account of violations of expected utility. At the aggregate level we observed support for the QALY model as we could not reject generalized marginality, the central condition of the QALY model. At the individual level there is much less support for the QALY model: a sizeable proportion of our subjects violated generalized marginality and the observed deviations were too large to be caused by elicitation and preference imprecisions alone.

We also tested for utility independence, a less restrictive preference condition than generalized marginality, which still implies a tractable model. Utility independence was supported at the aggregate level. At the individual level we found more support for utility independence than for generalized marginality. For a substantial proportion of our subjects the observed deviations from utility independence can reasonably be

attributed to the elicitation procedure and preference imprecision. Our aggregate findings on utility independence are consistent with the findings of Spencer and Robinson (2007) in spite of the differences in response mode and experimental design between their and our study. The data in Spencer and Robinson (2007) and in our study reinforce each other and provide a strong case for utility independence at the aggregate level. Spencer and Robinson do not report individual-level results.

Caveats—The decision tasks used in our experiment were cognitively demanding. Subjects had to take into consideration several dimensions simultaneously (probability, quality of life, and duration and sequence of the health states). We took several precautions to try and keep the experimental tasks as simple as possible by using just four easily imaginable color-coded health states, by using only three time periods of equal length, and by using a computer-run choice-based questionnaire. Nevertheless, subjects may have adopted simplifying heuristics to facilitate responding. Two such heuristics might a priori be particularly plausible.

First, subjects may have made the tasks easier by targeting towards probability 0.50. We had no indication that subjects indeed used this heuristic. First, most elicited probabilities differed significantly from 0.50. Second, there was no subject for whom all elicited indifference probabilities fell between 0.40 and 0.60. The data do not suggest that subjects were towards another probability either. When we compared differences in elicited probabilities across unrelated decision tasks (e.g. compare GM1-I with GM3-II) then many significant differences were observed. The latter observation also shows that the fact that we could not reject generalized marginality and utility independence at the aggregate level was not due to a lack of power in our study. Empirically meaningful differences in elicited probabilities can be elicited in our sample.

A second heuristic that subjects could have employed in the utility independence questions was to cancel out the common health states. For example, in the comparison between (p:LLM; MNM) and (p:LMM; MMM), task UI1-I, subjects may have made the task easier by eliminating the common health state M in the third period. Adopting such a strategy would make the two parts of each test of utility independence identical and would create artificial support for utility independence. The experimental design

took care to avoid that subjects would use this heuristic. In particular, we randomized the order of the tests so that subjects never faced the two parts of a test consecutively. It cannot be excluded though that at least some subjects adopted this cancellation heuristic in spite of the precautions we took.

We used students as subjects. We do not believe that this limits the generalizability of our findings. Empirical evidence on health utility measurement has shown that there exist no significant differences between the patterns of responses obtained from convenience samples and those obtained from representative samples from the general population. For a review see de Wit, van Busschbach, and de Charro (2000) and for a more recent comparison Bleichrodt, Doctor, and Stolk (2005).

Our results show that many subjects deviate from generalized marginality casting doubt on the descriptive appeal of the QALY model. These results say nothing about the normative validity of generalized marginality. One might argue that it is desirable for normative reasons to accept generalized marginality and interpret the deviations that we observed as irrationalities that reflect biases in time aggregation that we should seek to correct. We do not agree with this view. We do not consider generalized marginality normative and, as we explained in Section 2, there are good reasons why people may deviate from it.

We implicitly assumed that QALYs should reflect individual preferences for health. There is an alternative, extra-welfarist, strand in the literature, which takes QALYs as a measure of health and not necessarily as a reflection of people's preferences for health. The two approaches are not necessarily incompatible. Extra-welfarists use preference-based quality weights to quantify QALYs which suggests that even in the extra-welfarist approach individual preferences are important. Further, even if one takes the position that QALY are only a measure of health then one would expect that people's preferences are increasing in QALYs (better health is desirable). Given the uniqueness properties of utility, this essentially means that QALYs should reflect individual preferences. Hence, even in the extra-welfarist approach people's preferences and consequently our tests are important.

Our tests are robust to many violations of expected utility. As we explained in Section 2, the utility model that we assumed includes as special cases many of the theories of decision under risk that are available today. Hence, our results are not affected by the deviations from expected utility modeled by these theories. This is not to say that our results are robust to all deviations from expected utility. Our results depend on the validity of the general utility model and, even though very general, the model is not consistent with any preference pattern. For instance, the model makes no distinction between gains and losses and, hence, it is not robust to loss aversion. Developing tests that are robust to loss aversion is an important challenge. That said, our method corrects for more biases than any previous study and, hence, our tests are the most powerful tests of the QALY model available today.

Implications—Our results provide support for the QALY model at the aggregate level. It should be pointed out though that this conclusion is based on three tests only. It should also be kept in mind that we only used mild to moderate health states to avoid considerations like maximal endurable time. Our conclusions may no longer hold when more severe health states are involved. More evidence is needed and we invite other researchers to try and replicate our findings using other experimental designs.

At the individual level, the support for QALYs appears weaker. Our data suggest that QALYs cannot be applied in individual medical decision making without some additional tests of the decision maker's preference structure. The tests developed in this chapter may be helpful in doing so. In interpreting our results at the individual level, one should keep in mind though that the tasks were demanding and that there was a possibility of substantial imprecision in subjects' responses. Hence, this single study should not be taken as conclusive evidence of the validity of QALYs at the individual level.

Even when QALYs are found not to hold, not all is lost. Our results, suggest that there is more support for utility independence at the individual level. Utility independence still implies a tractable model that can be applied in practice. Hence, in contrast with a frequently voiced belief that QALYs are not consistent with people's preferences for

health, the overall message of this chapter is supportive of the use of QALY-type models in health economics.

Appendix A: Description of health states used in the experiment

KKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKK

Health state K

In a health state K, your health is characterized by:

<i>Mobility</i>	No problems walking about
<i>Self-Care</i>	No problems with self-care
<i>Usual Activities</i>	No problems with performing usual activities (for example, work, study, housework, family, leisure activities)
<i>Pain/Discomfort</i>	No pain or discomfort
<i>Anxiety/Depression</i>	Not anxious or depressed

LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL

Health state L

In a health state L, your health is characterized by:

<i>Mobility</i>	No problems walking about
<i>Self-Care</i>	No problems with self-care
<i>Usual Activities</i>	No problems with performing usual activities (for example, work, study, housework, family, leisure activities)
<i>Pain/Discomfort</i>	Moderate pain or discomfort
<i>Anxiety/Depression</i>	Not anxious or depressed

KKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKK

LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL

MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM

Health state M

In a health state M, your health is characterized by:

<i>Mobility</i>	No problems walking about
<i>Self-Care</i>	No problems with self-care
<i>Usual Activities</i>	No problems with performing usual activities (for example, work, study, housework, family, leisure activities)
<i>Pain/Discomfort</i>	Moderate pain or discomfort
<i>Anxiety/Depression</i>	Moderately anxious or depressed

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

Health state N

In a health state N, your health is characterized by:

<i>Mobility</i>	Some problems walking about
<i>Self-Care</i>	No problems with self-care
<i>Usual Activities</i>	Some problems with performing usual activities (for example, work, study, housework, family, leisure activities)
<i>Pain/Discomfort</i>	Moderate pain or discomfort
<i>Anxiety/Depression</i>	Moderately anxious or depressed

MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

Appendix B: Explanation of the bisection method

The bisection method used to generate the iterations is illustrated in Table A1 for task GM1I. The option that is chosen is printed in bold. The starting probability in the iterations was determined randomly. Depending on the choice made, the probability was increased or decreased. The size of change in the second iteration was half the difference between the probability in the first question and 0 or half the difference between the probability in the first question and 1. Which one was chosen depended on the subject's choice. The size of the change in the remaining iterations was half the size of the change in the previous question. The iteration process ended when the difference between the probability and the previous probability was less than 0.05. The iteration process resulted in an interval within which the indifference value should lie. The midpoint of this interval was taken as the indifference value. For example, in Table A1 the indifference value for p should lie between 0.63 and 0.68. Then we took as the indifference value 0.66.

Table A1: An illustration of the bisection method.

Iteration	Offered choices
1	(0.72:MMN; MNN) ~ (0.72:KMN; NNN)
2	(0.36:MMN; MNN) ~ (0.36:KMN; NNN)
3	(0.54:MMN; MNN) ~ (0.54:KMN; NNN)
4	(0.63:MMN; MNN) ~ (0.63:KMN; NNN)
5	(0.68:MMN; MNN) ~ (0.68:KMN; NNN)
Indifference value	0.66

References

- Bleichrodt, H., Abellan J. M., Pinto J. L., Mendez I. (2007). Resolving inconsistencies in utility measurement under risk: Tests of generalizations of expected utility. *Management Science* 53, 469-482.
- Bleichrodt, H., Doctor J. N., Stolk E. A. (2005). A nonparametric elicitation of the equity-efficiency trade-off in cost-utility analysis. *Journal of Health Economics* 24, 655-678.
- Bleichrodt, H., Gafni A. (1996). Time preference, the discounted utility model and health. *Journal of Health Economics* 15, 49-66.
- Bleichrodt, H., Johannesson M. (1997). The validity of QALYs: An empirical test of constant proportional tradeoff and utility independence. *Medical Decision Making* 17, 21-32.
- Bleichrodt, H., Pinto-Prades J.-L. (2006). Conceptual foundations for health utility measurement. In: Jones, A. M. (Eds.), *The Elgar companion to health economics*. Edward Elgar, Aldershot, 347-358.
- Bleichrodt, H., Quiggin J. (1997). Characterizing QALYs under a general rank dependent utility model. *Journal of Risk and Uncertainty* 15, 151-165.
- Brazier, J., Dolan P., Karampela K., Towers I. (2006). Does the whole equal the sum of the parts? Patient-assigned utility scores for ibs-related health states and profiles. *Health Economics* 15, 543-551.
- Butler, D., Loomes G. (2007). Imprecision as an account of the preference reversal phenomenon. *American Economic Review* 97, 277-297.
- de Wit, G. A., van Busschbach J. J., de Charro F. T. (2000). Sensitivity and perspective in the valuation of health status. *Health Economics* 9, 109-126.
- Doctor, J. N., Bleichrodt H., Lin J. H., forthcoming. Health utility bias: A meta-analytic evaluation. *Medical Decision Making*

- Doctor, J. N., Bleichrodt H., Miyamoto J., Temkin N. R., Dikmen S. (2004). A new and more robust test of QALYs. *Journal of Health Economics* 23, 353-367.
- Dolan, P. (1997). Modeling valuations for Euroqol health states. *Medical Care* 35, 1095-1108.
- Dubourg, W. R., Jones-Lee M. W., Loomes G. (1994). Imprecise preferences and the WTP-WTA disparity. *Journal of Risk and Uncertainty* 9, 115-133.
- Frederick, S., Loewenstein G. F., O'Donoghue T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature* 40, 351-401.
- Guerrero, A. M., Herrero C. (2005). A semi-separable utility function for health profiles. *Journal of Health Economics* 24, 33-54.
- Krabbe, P. F. M., Bonsel G. J. (1998). Sequence effects, health profiles, and the QALY model: In search of realistic modeling. *Medical Decision Making* 18, 178-186.
- Kuppermann, M., Shiboski S., Feeny D., Elkin E. P., Washington A. E. (1997). Can preference scores for discrete states be used to derive preference scores for entire paths of events? *Medical Decision Making* 17, 42-55.
- Luce, R. D. (2000). *Utility of gains and losses: Measurement-theoretical and experimental approaches*. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey.
- Mackeigan, L. D., O'Brien B. J., Oh P. I. (1999). Holistic versus composite preferences for lifetime treatment sequences for type 2 diabetes. *Medical Decision Making* 19, 113-121.
- Miyamoto, J. M. (1988). Generic utility theory: Measurement foundations and applications in multiattribute utility theory. *Journal of Mathematical Psychology* 32, 357-404.
- Miyamoto, J. M., Wakker P. P. (1996). Multiattribute utility theory without expected utility foundations. *Operations Research* 44, 313-326.

- O'Brien, B. J., Drummond M. F. (1994). Statistical versus quantitative significance in the socioeconomic evaluation of medicines. *Pharmacoeconomics* 5, 389-398.
- Quiggin, J. (1981). Risk perception and risk aversion among Australian farmers. *Australian Journal of Agricultural Economics* 25, 160-169.
- Richardson, J., Hall J., Salkeld G. (1996). The measurement of utility in multiphase health states. *International Journal of Technology Assessment in Health Care* 12, 151-162.
- Spencer, A. (2003). A test of the QALY model when health varies over time. *Social Science and Medicine* 57, 1697-1706.
- Spencer, A., Robinson A. (2007). Test of utility independence when health varies over time. *Journal of Health Economics* 26, 1003-1013.
- Stalmeier, P. F. M., Wakker P. P., Bezembinder T. G. G. (1997). Preference reversals: Violations of unidimensional procedure invariance. *Journal of Experimental Psychology: Human Perception and Performance* 23, 1196-1205.
- Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 28, 332-382.
- Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty* 32, 101-130.
- Treadwell, J. R. (1998). Tests of preferential independence in the QALY model. *Medical Decision Making* 18, 418-428.
- Tversky, A., Sattath S., Slovic P. (1988). Contingent weighting in judgment and choice. *Psychological Review* 95, 371-384.
- van der Pol, M. M., Cairns J. (2002). A comparison of the discounted utility model and hyperbolic discounting models in the case of social and private intertemporal preferences for health. *Journal of Economic Behavior and Organization* 49, 79-96.

van Osch, S. M. C., Wakker P. P., van den Hout W. B., Stiggelbout A. M. (2004).
Correcting biases in standard gamble and time tradeoff utilities. *Medical
Decision Making* 24, 511-517.

A Reply to Gandjour and Gafni

Summary

Gandjour and Gafni (2010) criticize our paper (Bleichrodt and Filko, 2008) on two counts. Their first point of criticism is ill-founded and results from many mathematical mistakes. The second is due to a lack of understanding of the general principles of empirical research.

Keywords: QALYs; Utility Theory

¹⁴ This chapter was published as Bleichrodt, H., Filko, M. (2010). A reply to Gandjour and Gafni. *Journal of Health Economics* 29, 329 – 331. We are grateful to Jason N. Doctor and Peter P. Wakker for helpful comments. Han Bleichrodt's research was made possible through a grant from the Netherlands Organization for Scientific Research (NWO). Martin Filko's research was supported by a grant from DSW.

In Bleichrodt and Filko (2008) we performed new tests of the QALY model when health varies over time. The novelty of our tests is that they control for violations of expected utility. It is well known that people do not behave according to expected utility (Starmer, 2000) and these violations may have confounded previous tests of the QALY model. Our experimental data supported QALYs at the aggregate level, but not at the individual level.

In a comment, Gandjour and Gafni (2009) criticize our paper on two grounds. First, they argue that it is possible that the condition we tested, *generalized marginality*, is not sufficient to imply the QALY model. In other words, subjects may simultaneously satisfy generalized marginality and violate the QALY model. Second, Gandjour and Gafni argue that we cannot make generalized statements about preferences because our sample is not representative. Related to this, they argue that we cannot conclude in support of a particular model based on a limited number of tests because the variety of health profiles is essentially endless. In this reply we will show that Gandjour and Gafni's first point of criticism is wrong. Their arguments contain many mathematical mistakes implying that their counterexamples are wrong and, therefore, that all their corresponding speculations are irrelevant. Their other points of criticism are completely standard (Popper, 1934, 1963) and reflect a lack of understanding of the general principles underlying all empirical studies in all fields of science. They also reflect poor reading as these points have actually been acknowledged and discussed in our paper.

4.1 First criticism: Support for generalized marginality and violations of the QALY model can coexist

There are three principal problems with Gandjour and Gafni's first point of criticism, which we will outline below.

First problem: Eq. (1) is ambiguous and ill-defined¹⁵

The first fundamental problem is that the model of Gandjour, (2008), which underlies Gandjour and Gafni's (2009) analysis and is stated in their Eq.(1), is not well-defined. A problem that recurs throughout their comment is that even though Gandjour and Gafni (2009) use mathematical derivations, they do not follow the rules and logic of mathematics (Suppes, 1957).

According to the left hand side of Eq.(1) u depends only on health states a, b , and c . However, on the right hand side of Eq.(1) the distributions $L(b)$ and $L(c)$ also appear. If these distributions play a role then they should also appear in the argument of the function. Then the utility of an outcome depends not only on the outcome itself, but on the whole distribution that it is part of. The model then loses all its tractability and becomes completely general without any predictive power. In particular, it is unclear how the formula should be applied when computing probability weighted averages such as in expected utility or its generalizations. Gandjour (2008) claims that expected utility should not be used to compute probability weighted averages but in Eq.(4) of their comment Gandjour and Gafni (2009) do use expected utility to compute probability weighted averages.

It is further a complete mystery where the functions L come from. Are these population statistics, marginal distributions or are they specific to the prospects that are being considered? Moreover, given the many parameters in Eq. (1) and their unclear nature identifiability of the model is also a problem.

There are two additional inaccuracies related to Eq.(1). Gandjour and Gafni (2009) call Eq.(1) additive, which it is not. A function is normally called additive if it is

additively decomposable, which implies strong separability. It is obvious that Eq. (1) does not satisfy strong separability. Apparently the authors use the term additive each time they discern an additive operation amidst other mathematical operations. We will ignore their claims about additivity in what follows.

A second inaccuracy is that Gandjour and Gafni use the same symbol u for several different things. In Eq.(1) u is used both as a function of a sequence of health states, as a function of the single-period health states, and as a function of the function L , representing the distribution of the health states within a period. This ambiguity about what the functions represent makes it hard to discuss the theory.

Because Gandjour's (2008) model is ill-defined and ambiguous, it is impossible to understand exactly what Gandjour and Gafni mean. Nevertheless we will try our best to interpret their writings as good as we can.

Second problem: Eq. (2) is wrong.

The second problem is that their claims made in Eq.(2) are unsubstantiated and wrong. Gandjour and Gafni claim that if strong separability, or additive utility independence as they call it, is imposed on top of Eq.(1) then Eq. (2) results. No proof is given for this claim and we will show that it is wrong. Gandjour and Gafni claim that additive separability implies that all λ 's must be equal to zero. Suppose, in contrast with Gandjour and Gafni's claim, that at least one of the λ 's is unequal to zero. Say $\lambda(a) = 1$. Suppose also that $u(L(b)) = u(L(c)) = 0$ for all b and c . Then Eq.(1) in Gandjour and Gafni becomes

$$u(a,b,c) = u(a) + u(b) + u(c) + u(b) + u(c) = u(a) + 2u(b) + 2u(c),$$

which is an additively decomposable form and which satisfies strong separability and additive utility independence. Hence, it is not true that strong separability or additive utility independence implies Eq.(2). This simple counterexample shows that Eq.(2) in Gandjour and Gafni (2009) is wrong, that their claims about the λ 's being equal to

¹⁵ Throughout this reply, the equation numbers refer to the equations in Gandjour and Gafni (2009).

zero are wrong, and that all the claims made later in the paper about generalized marginality and Eq.(2) are wrong.

Third problem: Eq.(4) and, hence, Gandjour and Gafni's counterexample against generalized marginality, is wrong.

We finally show that Eq.(4) in which Gandjour and Gafni (2009) derive what they believe generalized marginality tests is wrong. Before we do so, we must correct two mistakes in their Eq.(3), which describes our test of generalized marginality. A first problem with Eq.(3) is that Gandjour and Gafni, once again, violate the rules of logic and use different symbols to denote identical things. According to the rules of logic it is possible, for example, that a_I and a_{II} in Gandjour and Gafni's Eq.(3) are different. In the definition of generalized marginality they have to be identical. We will therefore ignore subscripts in what follows and simply write $a_I = a_{II} = a$, $c_I = c_{II} = c$ etc.

A second problem with Eq.(3) is that in the prospects on the right hand sides of the two indifference signs a'' appears twice. This is wrong. In each of these two prospects, the second term a'' has to be different from the first. We assume that this is a typo and that the authors had in mind to write a''' for the second terms.

Let us now explain the problems with Eq.(4). A first problem is that Gandjour and Gafni use expected utility. In our paper we use a much more general model than expected utility and Gandjour and Gafni should have shown that their conclusion holds under this more general model. A second problem is that Eq.(4) contains a term $L(b_I - b_{II})$. Why does b_{II} suddenly appear within brackets? This can only be if L , whatever it is, is linear (Aczel, 1966, Theorem 1, p.34). But such linearity has never been assumed. Moreover, b_{II} does not appear in Eq.(3) so where does it come from?

However, the fundamental problem with Eq.(4) is that it is wrong. Assuming Eq.(1) and expected utility as Gandjour and Gafni do, and following the same line of analysis as they do we obtain that the difference between the prospects on the left hand sides of the indifference signs is equal to:

$$p[u(b) - u(b'') + \lambda(a)(u(b) - u(b'')) + u(c)(\lambda(b) - \lambda(b'')) + \lambda(a)(u(L(b)) - u(L(b'')))] + u(L(c))(\lambda(b) - \lambda(b''))] + (1-p)[u(b') - u(b''') + \lambda(a')(u(b') - u(b''')) + u(c)(\lambda(b') - \lambda(b''')) + \lambda(a')(u(L(b')) - u(L(b''')))] + u(L(c))(\lambda(b') - \lambda(b'''))].$$

And the difference between the prospects on the right hand sides of the indifference signs is equal to:

$$p[u(b) - u(b'') + \lambda(a'')(u(b) - u(b'')) + u(c)(\lambda(b) - \lambda(b'')) + \lambda(a'')(u(L(b)) - u(L(b'')))] + u(L(c))(\lambda(b) - \lambda(b''))] + (1-p)[u(b') - u(b''') + \lambda(a''')(u(b') - u(b''')) + u(c)(\lambda(b') - \lambda(b''')) + \lambda(a''')(u(L(b')) - u(L(b''')))] + u(L(c))(\lambda(b') - \lambda(b'''))].$$

Deleting common terms this implies that generalized marginality tests whether

$$p[\lambda(a)(u(b) - u(b'') + u(L(b)) - u(L(b'')))] + (1-p)[\lambda(a')(u(b') - u(b''') + u(L(b')) - u(L(b''')))]$$

=

$$p[\lambda(a'')(u(b) - u(b'') + (u(L(b)) - u(L(b''))))] + (1-p)[\lambda(a''')(u(b') - u(b''') + u(L(b')) - u(L(b''')))].$$

This is clearly different from what Gandjour and Gafni obtain. Contrary to what Gandjour and Gafni claim the terms involving $L(\cdot)$ do not cancel. Having shown that Gandjour and Gafni's derivations and, hence, their counterexample, are wrong, all their speculations that follow Eq.(4) become irrelevant and we can safely ignore them.

4.2 Second Criticism: No General Statements are Possible

Regarding their second point of criticism we can be short: their expressed concerns are completely standard and are actually acknowledged in our paper.

The issue of representativeness is discussed in the third paragraph on page 1247. It is common to use convenience samples such as students to first test new decision

concepts. Consider, for example, Kahneman and Tversky (1979), the second most cited paper in economics since 1970 (Kim, Morse, and Zingales, 2006), which introduced prospect theory, the theory for which Kahneman was awarded the Nobel prize in economics in 2002. Kahneman and Tversky (1979) is entirely based on the responses of students and university faculty. Later studies then tested these new concepts in more general samples. In our case the new decision principle was generalized marginality. Because the test was new it made sense to first employ a convenience sample. Future studies should try to replicate our findings in more general samples.

Regarding the limited number of tests, it is well-known @ (Popper, 1934, 1963) that a hypothesis can never be proved right and can only be shown to be false. The classical example is the hypothesis “all swans are white.” This hypothesis can never be proved right but can be falsified by observing one single black swan. Gandjour and Gafni have nothing new to add here. According to Popper data that are in line with the theory “corroborate” the theory. Our general conclusion, repeated below, is entirely consistent with Popper (1934, 1963):

“Our results provide support for the QALY model at the aggregate level. It should be pointed out though that this conclusion is based on three tests only. It should also be kept in mind that we only used mild to moderate health states to avoid considerations like maximal endurable time. Our conclusions may no longer hold when more severe health states are involved. More evidence is needed and we invite other researchers to try and replicate our findings using other experimental designs.” (p.1247)

Let us end by correcting one final mistake in Gandjour and Gafni’s (2009) comment. They imply that we cite Spencer and Robinson (2007) as providing support for generalized marginality at the aggregate level. Once again, they did not read carefully. On page 1246 we wrote: “our aggregate findings on *utility independence* [emphasis added] are consistent with the findings of Spencer and Robinson (2007).” As we point out in our paper, if utility independence holds but generalized marginality is violated then period-specific utilities can still be defined and utility remains tractable.

4.3 Conclusion

Gandjour and Gafni (2009) criticize our paper on two counts. Their first point of criticism is ill-founded and results from many mathematical mistakes that they make. The second is due to a lack of understanding of the general principles of empirical research.

References

- Aczel, J. (1966). Lectures on functional equations and their applications. Academic Press, New York.
- Bleichrodt, H., Filko M. (2008). New tests of QALYs when health varies over time. *Journal of Health Economics* 27, 1237-1249.
- Gandjour, A. (2008). Incorporating feelings related to the uncertainty about future health in utility measurement. *Health Economics* 17, 1207-1213.
- Gandjour, A., Gafni A. (2009). The additive utility assumption of the QALY model revisited.
- Kahneman, D., Tversky A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica* 47, 263-291.
- Kim, E. H., Morse A., Zingales L. (2006). What has mattered to economics since 1970. *Journal of Economic Perspectives* 20, 189-202.
- Popper, K. R. (1934). *Logik der forschung*. Julius Springer Verlag, Vienna.
- Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. Routledge & Kegan Paul, London.
- Spencer, A., Robinson A. (2007). Test of utility independence when health varies over time. *Journal of Health Economics* 26, 1003-1013.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 28, 332-382.

Suppes, P. (1957). *Introduction to logic*. Van Nostrand, New York.

Making Case-Based Decision Theory Directly Observable

Summary

Gilboa & Schmeidler's case-based decision theory (CBDT) is an alternative to Savage's state-space model for uncertainty. Preferences are determined by similarities with cases in memory. This chapter introduces a nonparametric method to elicit CBDT, requiring no commitment to parametric families and relating directly to decisions. An experiment on real estate investments demonstrates the feasibility of our method. Our implementation of real incentives avoids not only the income effect, but also interaction between different memories. We confirm CBDT's predictions with however one violation of separability of cases in memory. CBDT gives plausible predictions and new insights into (real estate investment) decisions.

Keywords: case-based decision theory, similarity weights, random incentive system, nonexpected utility, ambiguity, real estate investments

Case-based decision theory (CBDT) was introduced as an alternative to Savage's (1954) classical state-space model for decision under uncertainty. In Savage's model, acts map states to outcomes and preferences are expressed over acts. Savage's model is commonly used today and forms the basis for generalizations of expected utility that incorporate ambiguity, especially in the version of Anscombe & Aumann (1963).

In CBDT, preferences are determined by cases in the decision maker's memory and their similarity with the decision problem at hand. CBDT has several advantages over the classical model. No counterfactual events or outcomes need to be considered (Gilboa & Schmeidler 2001, henceforth GS, pp. 43, 93-95), and CBDT naturally fits with our everyday thinking.¹⁶ The primary achievement of GS was to connect case-based models of information processing, widely used in artificial intelligence (Aha, Marling, & Watson 2005; Hüllermeier 2007; Riesbeck & Schank 1989) and other fields (Dubois et al. 1999; Greco, Matarazzo, & Slowinski 2008; Hertwig et al. 2004; Stewart, Chater, & Brown 2006), with economic decision making.

Gilboa & Schmeidler (1995, and many followups) and Eichberger & Guerdjikova (2011) provided preference foundations of CBDT, demonstrating its theoretical soundness. There is, however, a dearth of empirical applications and those that exist have as yet focused on direct introspective judgments (Lovallo, Clarke, & Camerer 2012) or on parametric fittings (Gayer, Gilboa, & Lieberman 2007; Golosnoy & Okhrin 2008; see also Guerdjikova 2008 p. 112, and Pape & Kurtz 2012). Such fittings require a commitment to particular parametric families of similarity weights, usually based on a particular distance function, and to particular error theories. These commitments introduce distortions when the parametric assumptions do not correspond with people's preferences. They are extra problematic for new concepts such as CBDT's similarity weights, because we do not know much about their properties (Guerdjikova 2008; Pape & Kurtz 2012). Similarity judgments may even

¹⁶ Greenspan (2004, p. 38) wrote: "how ... the economy might respond to a monetary policy initiative may need to be drawn from evidence about past behavior during a period only roughly comparable to the current situation." Charness & Levin (2005) consider situations where naïve case-based reasoning leads to opposite, inferior, decisions than Bayesian reasoning, but still find that about half of their subjects follow the case-based reasoning.

violate basic properties of distance measures (Grosskopf, Sarin, & Watson 2008 §1.2).

We will introduce a nonparametric method to measure similarity weights that avoids all parametric assumptions. It is a close analog of de Finetti's (1931) betting odds system for measuring subjective probabilities, a well-known nonparametric measurement method in classical decision theory.¹⁷ Nonparametric measurements provide correct results in full generality, revealing the right properties whatever they are. The required measurements are elementary and can be carried out using only paper and pencil. They immediately reveal the empirical meaning of the concepts measured in terms of preferences without complex computer fitting intervening. They can be used in prescriptive and interactive sessions to determine optimal values of parameters (Keeney & Raiffa 1976). They are also useful in preference axiomatizations, which often amount to excluding inconsistencies in nonparametric measurements (Wakker 2010 p. 8).

We demonstrate the feasibility of our measurement technique in two experiments on real estate investment, a domain that is particularly prone to case-based reasoning (Gayer, Gilboa, & Lieberman 2007). A complication in experimentally testing CBDT is that it not only entails a deviation from Savage's uncertainty model but, more fundamentally, from the classical revealed preference paradigm: It varies information in memory rather than available choice alternatives. Consequently, we need a new mechanism to implement incentives in experiments. The popular random incentive system serves to avoid income effects (interactions between different outcomes received), but for CBDT, we additionally have to avoid interactions between the different memories that are used in the different decision problems. A complication in designing such an incentive mechanism is that people cannot deliberately forget information. Hence we developed an adaptation of the random incentive system to CBDT, which is explained in Section 4.

Our experimental findings confirmed most of the predictions of CBDT. Yet we did find some evidence of interactions between different cases in a memory. This

¹⁷ The decision-theoretic term nonparametric should not be confused with the statistical term, as discussed by Gilboa, Lieberman, & Schmeidler (2011) in the context of CBDT.

suggests that CBDT's assumption that cases are separable is too restrictive and that generalizations of this assumption are desirable (Eichberger & Guerdjikova 2011; Peski 2011). Our findings on real estate investments are plausible. They underscore that CBDT is a viable alternative to classical revealed preference for obtaining new insights into (real estate investment) decisions.

5.1 Case-based decision theory versus classical revealed preference: varying memory instead of available choice options

In classical decision theory, choices are usually derived from other decisions (Gilboa, Schmeidler, & Wakker 2002, GSW henceforth, beginning of §3; Jahnke, Chwolka, & Simons 2005 pp. 20-21; Manski 2011; Starmer & Sugden 1991). In the revealed preference approach these other decisions concern variations in the set of available choice options. For example, in consumer demand theory, choice options are commodity bundles and sets of available options are budget sets. Usually, choices maximize a preference relation that describes all choices from pairs, and this preference relation is used as primitive. Savage's (1954) decision uncertainty model is one of many examples. Exceptions notwithstanding,¹⁸ variations in the set of available choice options are the almost exclusively used tool in decision theory today.

CBDT entails a fundamental departure from classical decision theory. Instead of varying the available choice options, CBDT varies the information available, termed memory (GS §4.2 and p. 94; GSW p. 485). A *memory* M contains cases. A *case* is a triple $(p,a,r) \in P \times A \times R$, where $p \in P$ is a *problem* encountered in the past, $a \in A$ is the *act* chosen there, and $r \in R$ is the *outcome* that resulted. The pair (p,a) is a *circumstance*. If a particular act a is chosen in a given problem, there is no need to specify which other acts were available (they were all less preferred than a , obviously.) In contrast with the classical paradigm (Savage 1954), there is no need to specify counterfactual events and/or outcomes in CBDT. For discussions about the

¹⁸ One exception is social choice theory (Arrow 1951) where the set of choice options is kept fixed and instead the preferences of the individuals in society vary.

weaknesses of the classical paradigm see Aumann 1971; Gilboa 2009; GS; Karni & Vierø 2011; Luce 2000 §1.1.6.1.

CBDT assumes a *utility function* u mapping outcomes to the reals, and a nonnegative *similarity function* $s((p,a),(q,b))$. The latter function, which is the new subjective parameter introduced by CBDT, describes for each pair of circumstances (p,a) and (q,b) how similar they are. We assume that the actual problem faced is p , that an act (= choice option) a is to be chosen from a set D of available acts, and that the act a chosen is the one maximizing

$$U(a) = \sum_{(q,b,r) \in M} s((p,a),(q,b))u(r). \quad (1.1)$$

The similarity weights determine the exchange rate between utility units under different circumstances. Under CBDT, preferences depend on the memory M . Hence we denote the preference relation over acts by \succsim_M . Strict preference and indifference are denoted \succ_M and \sim_M , respectively.

We use a general version of CBDT with act similarity (GS p. 51). Other versions are discussed in the discussion section. The following example illustrates the procedures of CBDT, and its most critical condition, separability of cases (GS p. 66, A2, combination).

EXAMPLE 1.1 [separability of different cases]. Assume that:

- (1) Dish (act) a is chosen from menu $D = \{a,b\}$ if the agent's memory contains the following four cases: A choice of dish a' during the last three Fridays gave a moderately positive outcome each time. A choice of dish b' on Friday a month ago gave a very positive outcome.
- (2) Dish a is chosen from $D = \{a,b\}$ if the agent's memory contains the following four cases: A choice of dish a' during the last two Tuesdays gave a moderately positive outcome each time. A choice of dish b' the two Tuesdays before did so too.

Then

- (3) Dish a is also chosen from $D = \{a,b\}$ if the agent's memory contains both the four cases in (1) and the four cases (2).

The example illustrates the departure of CBDT from the classical paradigm. The set of choice alternatives to choose from, $D = \{a,b\}$, is kept fixed and for the cases in memory the available acts are not even specified. Hence, variations in the set of available choice options is not used. What varies is the information upon which the actual choice is to be based. All outcomes considered have really been experienced and, hence, no use is made of hypothetical situations. \square

Variations in available information and experience are often central when we make decisions under uncertainty. Gilboa and Schmeidler (2001) provide many examples. Hence, CBDT primarily serves as a paradigm that is alternative to the classical paradigm for uncertainty of Savage (1954) and Anscombe & Aumann (1963).¹⁹

The following uniqueness result holds for CBDT.

OBSERVATION 1.2. In Eq. 1.1, preferences are not affected if

- i. u is multiplied by a positive factor²⁰;
- ii. circumstance-dependent constants²¹ are added to the similarity weights;
- iii. all similarity weights are multiplied by a common positive factor. \square

¹⁹ The new paradigm of CBDT can also serve as a new approach for consumer theory (Gilboa & Schmeidler 1997b), adaptive optimization (Gilboa & Schmeidler 1996; Golosnoy & Okhrin 2008; Guerdjikova 2008; Jahnke, Chwolka, & Simons 2005), social structures (Blonski 1999), and inductive reasoning and probability assessment (Eichberger & Guerdjikova 2010; Gayer 2010; Gilboa, Lieberman, & Schmeidler 2010).

²⁰ See GSW, Theorem 2.1.

²¹ See GSW, Theorem 2.1, or Gilboa & Schmeidler (1997a the Theorem). Adding the same constant $c(q,b)$ to all $s((p,a),(q,b))$ increases the utility of all acts by the same constant $c(q,b)u(r)$, which does not affect preference. Section 3 gives a derivation when there are two acts in D . The constants added cannot depend on the acts a in D for otherwise preference would be affected.

Under minimal richness assumptions, only the invariance described in Observation 1.2 is permitted. Then the similarity weights are a kind of joint interval (cardinal) scales²² and u is a ratio scale. Thus, in general, the outcome with utility 0 is empirically meaningful and cannot be chosen arbitrarily. This follows because the sum of the similarity weights may differ from 1 or another constant and can depend on the memory under consideration (GS pp. 40, 43).

The outcome with utility 0 is called *neutral*. As we will explain in the next section, in our method we need not know the neutral outcome to be able to measure the similarity weights. We achieve this by a design where the terms corresponding to the unknown neutral outcome drop from the equations. Further discussions of neutral outcomes include GS (pp. 133, 148 ff.) and many other papers.

5.2 Direct (nonparametric) measurements of utility and similarity weights:

Theory

This section presents a theoretical analysis showing how the parameters of CBDT, utility and the similarity weights, can be measured nonparametrically. The basic procedure for similarity weights is the analog of de Finetti's betting odds system for measuring subjective probabilities. In short, if in a given situation improving a result for circumstance 1 by 5 leads to indifference, and improving a result for circumstance 2 by 3 also leads to indifference, then the proportion of the similarity weights of these two circumstances is 3:5. We now explain the procedure in more detail.

It suffices to consider only two acts to choose from. We therefore focus on this simple case and assume $D = \{a_0, a_1\}$ for a problem p . Here, and in what follows, we use notation that will be convenient for the experiment reported later. For each circumstance (q_j, b_j) , we define the difference between the similarity weights of the two acts considered:

²² See Guerdjikova (2008 pp. 109-110). In many applications of CBDT, further scaling conventions are imposed on the similarity weights that determine their 0 level, such as being 0 whenever the acts involved in the two circumstances are different (Blonski 1999; Gilboa & Schmeidler 1995 Theorems 1, 2; Gilboa & Schmeidler 1996, 1997a; Jahnke, Chwolka, & Simons 2005 pp. 17, 23). Then the

$$d_j = s((p, a_1), (q_j, b_j)) - s((p, a_0), (q_j, b_j)). \quad (2.1)$$

We call d_j the *decision weight* of the corresponding circumstance. Eq. 1.1 leads to the following decision criterion:

$$a_1 \underset{<}{\succ} a_0 \Leftrightarrow \sum_{(q_j, b_j, r_j) \in M} d_j u(r_j) \underset{<}{\geq} 0. \quad (2.2)$$

We call circumstance (q_j, b_j) *favorable* if $d_j > 0$, *neutral* if $d_j = 0$, and *unfavorable* if $d_j < 0$. Throughout this chapter it is understood that these terms are for a_1 versus a_0 (GSW p. 487). Eq. 2.2 illustrates the uniqueness results of Observation 1.2: We cannot observe more of similarity weights than the generated decision weights, and, further, decision weights are unique up to a common positive factor. This invariance is similar to ratio scales. A difference is that decision weights can be of either sign, whereas common ratio scales are usually only positive. Decision weights can also be zero.

We can also observe the signs of decision weights. That is, we can observe whether improving a result of a case changes preference favorably or unfavorably. Hence, for a given memory we can distinguish the pair of decision weights 1, 2 from the pair -1 , -2 even though they have the same ratio. This aspect complicates the mathematical and statistical analysis of similarity weights, in the same way as it complicates the analysis of sign-dependent ratios in general (Koerkamp et al. 2007).

GSW present two ways to measure utility that we discuss in Appendix A. The central topic of our chapter is the measurement of similarity weights. We assume that utilities are known. They may have been obtained using Appendix A, or they may be assumed linear, as in our experiment. We take a pair of default outcomes (r_0, r_1) for which the acts a_1 and a_0 are not indifferent and then consider two changes $r_0 \rightarrow r_0^2$ and $r_1 \rightarrow r_1^1$, each of which leads to indifference between a_1 and a_0 . Gilboa and Schmeidler have usually analyzed CBDT under the assumption that such outcomes r_0^2 and r_1^1 exist. To ensure their existence they used various diversity axioms or solvability/continuity

similarity weights are ratio scales. The requirement that similarity weights are nonnegative restricts the constants that can be added.

axioms (for the latter, see Gilboa & Schmeidler 1995 pp. 635-636 Axiom A2', or GSW p. 489: solvability). In the following theorem we also assume that r_0^2 and r_1^1 exist. The results of the Theorem are stated in theoretical terms that are not directly observable, but they will provide the groundwork for the observable results presented later.

THEOREM 2.1. Assume $D = \{a_0, a_1\}$ for problem p , and memories $M = M_a \cup M_p$ with $M_a = \{(q_0, c_0, r_0), (q_1, c_1, r_1)\}$. We denote $M = (r_0, r_1)$ because all else is fixed here.

Assume

$$a_0 \not\sim_{(r_0, r_1)} a_1, a_0 \sim_{(r_0^2, r_1)} a_1, \text{ and } a_0 \sim_{(r_0, r_1^1)} a_1 \quad (2.3)$$

for some r_0^2 and r_1^1 . Then

$$\frac{d_1}{d_0} = \frac{u(r_0^2) - u(r_0)}{u(r_1^1) - u(r_1)} \quad (2.4)$$

with both ratios well-defined and nonzero. Regarding the sign of a decision weight, i.e. the favorability of the corresponding circumstance, each line below gives a triple of equivalent statements:²³

(1): $d_0 > 0$; (2): $[u(r_0') > u(r_0^2) \Rightarrow a_1 >_{(r_0', r_1)} a_0]$; (3): $[u(r_0') < u(r_0^2) \Rightarrow a_1 <_{(r_0', r_1)}$
 $a_0]$;

(1): $d_0 < 0$; (2): $[u(r_0') > u(r_0^2) \Rightarrow a_1 <_{(r_0', r_1)} a_0]$; (3): $[u(r_0') < u(r_0^2) \Rightarrow a_1 >_{(r_0', r_1)}$
 $a_0]$;

(1): $d_1 > 0$; (2): $[u(r_1') > u(r_1^1) \Rightarrow a_1 >_{(r_0, r_1')} a_0]$; (3): $[u(r_1') < u(r_1^1) \Rightarrow a_1 <_{(r_0, r_1')}$
 $a_0]$;

(1): $d_1 < 0$; (2): $[u(r_1') > u(r_1^1) \Rightarrow a_1 <_{(r_0, r_1')} a_0]$; (3): $[u(r_1') < u(r_1^1) \Rightarrow a_1 >_{(r_0, r_1')}$
 $a_0]$. \square

The proof of Theorem 2.1 is in Appendix B. The theorem shows that, starting from memory (r_0, r_1) , a change from $u(r_0)$ to $u(r_0^2)$, which is weighted by d_0 , has the same nonzero effect as a change from $u(r_1)$ to $u(r_1^1)$ weighted by d_1 . This leads to Eq. 2.4.

Given that we can observe utility (Eqs. A.2 and A.3), Theorem 2.1 gives all the information that can be obtained about d_0 and d_1 and the underlying similarity weights. This follows from Observation 1.2, the subsequent discussion there, and the discussion following Eq. 2.2. The theorem shows that only the degree to which cases in memory are more favorable for a_1 than for a_0 is relevant for the choice between acts. To derive Eq. 2.4, we need not know which outcome is neutral, i.e. what the level of utility is, because this level drops from the equation.

We will assume linear utilities, which is reasonable for the moderate amounts used in the experiment. This way we avoid having to measure utility (see Appendix A), thus reducing the number of tasks that the subjects had to perform. We included several tests of linear utility and these supported our assumption. Linear utility was also assumed by Gilboa and Schmeidler in several papers (Gilboa & Schmeidler 1995 p. 613 and Axiom A4, and p. 635 and Axiom A4'; Gilboa & Schmeidler 1997a p. 50 and Axiom A3). For linear utility we obtain from Eq. 2.4:

$$\frac{d_1}{d_0} = \frac{r_0^2 - r_0}{r_1^1 - r_1} . \quad (2.5)$$

In our experiment we used negative weights d_0 . We therefore normalized Eq. 2.5. by division by $|d_0| = -d_0$, so that larger d_1 weights correspond with larger normalized weights. We then have

$$\frac{d_1}{-d_0} = \frac{r_0 - r_0^2}{r_1^1 - r_1} . \quad (2.6)$$

²³ We added primes in r_0^1 and r_1^1 below to distinguish them from r_0 and r_1 . The latter variables usually are no free variables but they play particular roles, such as true values in our experiments.

Eq. 2.6. forms the basis of the measurements reported below.

5.3 A CBDT version of the random incentive system

The implementation of incentives in experiments on CBDT raises a subtle issue, not present in traditional decision experiments based on classical revealed preference. In traditional experiments, subjects are typically asked to make many choices to obtain as much information as possible within the constraints of the experiments. Yet each of these choices has to be taken in isolation to avoid income effects due to interactions between outcomes.²⁴ Illuminating early discussions are in Savage (1954 p. 29) and Ramsey (1931 pp. 169-174). Hence the random incentive system (RIS) was introduced to implement incentive compatibility.²⁵ Subjects are asked to make many choices, but only one randomly selected choice is played out for real. Under some plausible assumptions, it is in the subjects' best interest to take each choice as isolated (Holt 1986; Starmer & Sugden 1991).

To measure CBDT, we have to consider choices under several memories. As in the classical RIS, we can play out one randomly selected choice for real at the end of the experiment, avoiding income effects. However, this procedure only works if in each choice the new memory *replaces* previous memories, rather than being added to them. Simply asking subjects to forget or ignore the information provided at previous choices may be possible in hypothetical choices with cooperative subjects (Gilboa & Schmeidler 1995 p. 621 points I and II), but it is impossible with real incentives and self-interested subjects.

To our knowledge, the only study that succeeded in implementing real decisions for CBDT is Grosskopf, Sarin, & Watson (2008). They used abstract cases so as to avoid prior memory effects, and told subjects that the different decisions (which were all implemented for real and income effects were therefore not excluded) did not affect

²⁴ If repeated purchases and multiple consumptions are relevant for the actual decision problem, then they have to be explicitly modelled that way. Then a choice option should describe combinations of purchases and of consumption bundles. A choice option is by definition what is to be chosen once.

²⁵ The first proposal that we are aware of is Savage (1954 p. 29).

each other and were independent. Subjects may however have figured that the randomizations all came from the same experimental implementation, using the same payoff function, and may still have perceived relations and similarities. We therefore opted for a different and more radical way to separate information in different decision situations, by developing a case-based version of the RIS. We also excluded income effects.

We told subjects that they would be asked many decisions, each with a different piece of information (= memory). One of these pieces was true. The others were made up by us (possibly still partly true). This was apparent in the experiment because different memories were usually mutually incompatible. Subjects did not know what the true information was. At the conclusion of the experiment, the decision implemented for real was based on the true information. It was explained to the subjects that to have their best choice implemented in the actual decision implemented at the end (and based on the true information), they should take every memory provided at a choice as the only true one for that choice. The memories provided before and after were irrelevant for the choice considered. Rational subjects should follow this advice. It is possible that subjects still thought that the information that was made up provided clues about the true information, or that subjects perceived a meta-lottery over pieces of information. These risks are similar to those in the traditional RIS. Isolated processing is in a subject's best interest in our design as it is in the traditional RIS.

5.4 Experiment 1 to measure similarity weights and to test CBDT

This section shows how we implemented the measurement of decision weights in our first experiment.

5.4.1 Stimuli

Our stimuli were based on the development of the prices of real estate in different provinces (states) of the Netherlands. Figure 4.1 depicts a map of the Netherlands with the different provinces indicated. In what follows, we will use a notation that makes the text easily accessible to readers unfamiliar with the Dutch geography so that knowledge of Figure 4.1 is not needed.

Subjects had to choose between two acts, a_0 and a_1 . Act a_0 yielded

$$€3 + 10x \tag{4.1}$$







if the price of a single-family house in the province of South-Holland increased by $x\%$ in the month after the experiment. The other act, a_1 , yielded the same payoff, but with x now referring to the percentage increase of an apartment in the Dutch province of North-Brabant. Figure 4.2 give two screenshots from the experiment.

FIGURE 4.1: Map of the Netherlands







FIGURE 4.2. Stimuli in experiment: three cases in memory

Here is the information about the value appreciation of different real estate types:

6 %			House in Limburg with appreciation of 6 %
15.9 %			House in Noord Holland with appreciation of 15.9 %
14.7 %			Apartment in Utrecht with appreciation of 14.7%

If annual appreciation for different houses and apartments in the past were according to this table, would you gamble on a house in Zuid Holland or on an apartment in Brabant?

You want to gamble on
(click on the picture)

Apartment in Noord Brabant			House in Zuid Holland		
-----------------------------------	---	---	------------------------------	---	---

Back

The cases in memory that we manipulated experimentally all concern one of the following three acts, being past gambles on related but different types of dwelling in different provinces in the Netherlands:

c_0 (house in North-Holland);

c_1 (house in Limburg);

c_2 (apartment in Utrecht).

We have numbered cases according to our prior expectation of their favorability for a_1 against a_0 , with c_2 most favorable and c_0 least favorable (which was later confirmed by the experiment). Thus readers need not know the Dutch conditions.. In particular, a_0 and c_0 will be perceived as very similar, and favorable outcomes under c_0 will enhance a preference for a_0 over a_1 .²⁶

Our subjects probably have more information than the information provided in the experiment: they entered the lab with some prior knowledge about real estate prices. We call this information their *prior memory*, and it is modeled through extra cases $M_p = \{q_j, b_j, r_j\}_{j>4}$, which are assumed subject-dependent and unobservable for us.²⁷ The *added memory* refers to the set of additional cases, which was manipulated in the experiment, and is denoted M_a . The complete memory is $M = M_a \cup M_p$. Because M_p is fixed, we can ignore it and equate the memory with M_a . The terms related to M_p will always drop from the equations in what follows.²⁸ This implies that biases in retrieving cases from memory such as the availability heuristic (Lovalló, Clarke, & Camerer 2012), do not affect our measurements.

Because each act c_j in memory corresponds with a different problem q_j in our application, we will usually refer to circumstances by only denoting the relevant acts in memory. The latter uniquely determining the corresponding problems. That is, problems are suppressed and circumstances are identified with acts in memory (Guerdjikova 2008 p. 109). From now on we use the term *act* only for a_0 and a_1 , the

²⁶ Our subjects, like almost everyone else, are unlikely to be finance specialists believing in a reversal (good past performance implies bad future performance) for real estate prices.

²⁷ We use indexes $j > 4$ because the indexes $j = 3, 4$ will be used in Experiment 2.

acts to be chosen from, and the term *circumstance* for c_0 - c_2 in memory. This is also why we use the symbol c in c_0 - c_2 . We define decision weights d_j as in Eq. 2.1. Thus, d_j indicates to what extent a good result for c_j supports the choice of a_1 rather than of a_0 . As explained, we expected that $d_2 > d_1 > d_0$; d_0 will usually be negative. Because we assume that utility is linear, we have $u(r_j) = r - c$, with c the subject-dependent and unobserved neutral outcome. As can be seen from Eq. 2.6. on which our measurements are based, we need not know the neutral outcome because it drops from the equation. A plausible neutral outcome in our experiment may be the risk-free interest rate (Golosnoy & Okhrin 2008).

The real outcomes r_0 , r_1 , and r_2 of the three circumstances c_0 , c_1 , and c_2 were 0.159, 0.06, and 0.147, respectively. These values are depicted in Figure 4.2. In other words, the annual rise in the price of a house in South-Holland over the past 3 years was 15.9%, it was 6% for a house in Limburg, and it was 14.7% for an apartment in Utrecht. These real values played a role similar as r_0 and r_1 in Theorem 2.1. That is, we considered departures from these values that produced indifferences between the acts a_0 and a_1 .

The indifferences were determined through an iteration process illustrated in Table 4.1. The table displays the answers of one of the subjects (Subject 35) in the experiment. The first row shows that we first asked the subject to choose between a_0 and a_1 for $r_0 = 15.9\%$ and $r_1 = 100\%$.²⁹ Not surprisingly, the subject chose a_1 . We then decreased r_1 to 0% and now he chose a_0 . By varying r_1 depending on the subject's choices we zoomed in on the value of r_1 for which the subject was indifferent between a_0 and a_1 . The recorded indifference value was the midpoint between the two outcomes where preference switched. In the table it is 19.8% (the midpoint between 19.1% and 20.6%).

²⁸ It is very unlikely that one of the cases in the added memory was already present in the prior memory. Hence we assume that M_a and M_p are disjoint.

²⁹ That is we asked them to choose between a house in South-Holland and an apartment in North-Brabant when it was given that the annual increase in house prices in North-Holland over the past 3 years was 15.9% and the annual increase in house prices in Limburg was 100% over the past 3 years.

TABLE 4.1

Iteration	r_0	r_1	Choice
1	15.9%	100%	a_1
2	15.9%	0%	a_0
3	15.9%	6%	a_0
4	15.9%	53%	a_1
5	15.9%	29.5%	a_1
6	15.9%	17.7%	a_0
7	15.9%	23.6%	a_1
8	15.9%	20.6%	a_1
9	15.9%	19.1%	a_0

Table 4.2 gives an overview of the questions that we asked in Experiment 1. There were 8 questions and the table entries denote the stimuli that were included. An empty cell means that that circumstance was not present in a memory. For example, we can see from Table 4.2 that circumstance c_2 (apartment in Utrecht) was not included in the second question. The outcomes printed in bold are those that were varied to produce indifference between a_0 and a_1 . The other outcomes, those not in bold, were always equal to their real values. The choice process shown in Table 4.1 corresponds to the second question.

The true value of the bold outcome was always presented as one of the choices. Hence, there was always one choice situation in which all outcomes were true, but subjects did not know in which situation this happened. For example, the choice situation for M^6 with real values is in Figure 4.2.

TABLE 4.2

	M^1	M^2	M^3	M^4	M^5	M^6	M^7	M^8
c_0	r_0	$\mathbf{r_0^2}$	r_0^{2+5}	$\mathbf{r_0^4}$	r_0	r_0	$\mathbf{r_0^7}$	r_0
c_1	$\mathbf{r_1^1}$	r_1	$\mathbf{r_1^3}$			$\mathbf{r_1^6}$	r_1	r_1
c_2				r_2	$\mathbf{r_2^5}$	r_2	r_2	$\mathbf{r_2^8}$

Each memory $M_a = M^j$ is specified by the outcomes below it, referring to the circumstances in the corresponding row. Thus $M^1 = \{(c_0, r_0), (c_1, r_1^1)\}$ and $M^7 = \{(c_0, r_0^7), (c_1, r_1), (c_2, r_2)\}$. In each case except c_0 at M^3 , the outcomes not in bold were the real outcomes. The bold outcome was varied to produce indifference between a_0 and a_1 .

5.4.2 Similarity weights in our design

We saw before that Subject 35 was indifferent between a_0 and a_1 for $r_1^1 = 19.8\%$ in the first choice question. For this value of r_1^1 the information for or against a_1 relative to a_0 provided by case (c_1, r_1^1) exactly offsets the information provided by (c_0, r_0) joint with M_p . The decision weight d_1 is positive for this subject because a_1 is preferred for larger values than r_1^1 and a_0 is preferred for smaller values than r_1^1 .

In general, the outcomes r_1^1 and r_0^2 are such that

$$a_1 \sim_{M_a \cup M_p} a_0 \text{ for } M_a = M^1 = \{(c_0, r_0), (c_1, r_1^1)\} \text{ and} \quad (4.2)$$

$$a_0 \sim_{M_a \cup M_p} a_1 \text{ for } M_a = M^2 = \{(c_0, r_0^2), (c_1, r_1)\}. \quad (4.3)$$

As another example, outcome r_2^8 is such that

$$a_0 \sim_{M_a \cup M_p} a_1 \text{ for } M_a = M^8 = \{(c_0, r_0), (c_1, r_1), (c_2, r_2^8)\}. \quad (4.4)$$

By Theorem 2.1 we have the following results, where following each statistic we

define a shorthand notation for it. For example, $\frac{d_1^1}{d_0^2}$ denotes the statistic estimating

$d_1/(-d_0)$ that can be derived from M^1 and M^2 , and $\frac{d_1^3}{d_0^2}$ denotes the statistic estimating

$d_1/(-d_0)$ that can be derived from M^1 and M^3 , with division by $-d_0$ explained in Eq. 2.6.³⁰

$$\frac{d_1}{-d_0} = \frac{r_0-r_0^2}{r_1^1-r_1} =: \frac{d_1^1}{d_0^2} (M^1 \& M^2). \quad (4.5)$$

$$\frac{d_1}{-d_0} = \frac{r_0-(r_0^2+5)}{r_1^1-r_1^3} =: \frac{d_1^3}{d_0^1} (M^1 \& M^3). \quad (4.6)$$

$$\frac{d_1}{-d_0} = \frac{r_0-r_0^7}{r_1^6-r_1} =: \frac{d_1^6}{d_0^7} (M^6 \& M^7). \quad (4.7)$$

$$\frac{d_2}{-d_0} = \frac{r_0-r_0^4}{r_2^5-r_2} =: \frac{d_2^5}{d_0^4} (M^4 \& M^5). \quad (4.8)$$

$$\frac{d_2}{-d_0} = \frac{r_0-r_0^7}{r_2^8-r_2} =: \frac{d_2^8}{d_0^7} (M^7 \& M^8). \quad (4.9)$$

By Theorem 2.1, our observations also reveal the signs of all d_j . Observation 1.2 then shows that our measurements reveal all the information that can be obtained about the similarity weights.

5.4.3 Sample and procedure

Subjects. $N = 53$ (26 female) undergraduate students from Erasmus University coming from diverse academic backgrounds signed up for the experiment. We decided not to drop any subject for erratic behavior (although we did treat zero decision weights as missing, as explained later in §4.5). Given the novelty of CBDT,

³⁰ Whenever the two relevant memories contain two cases, we can immediately apply Theorem 2.1. When both memories contain three cases (M^6 - M^8), there is one case that has the same outcome throughout; e.g., for M^6 and M^7 this concerns c_2 . We then apply Theorem 2.1 with this common case included in M_p . For example, to obtain Eq. 4.7, we take the M_p of Theorem 2.1 equal to our $M_p \cup \{(c_2, r_2)\}$.

it is not clear to what extent deviations from the model (such as nonmonotonicity in outcomes) can be interpreted as erratic or as a valid violation. Keeping all subjects increases the noise in our data and makes our tests conservative.

Stimuli. Table 4.2 describes the choices faced by the subjects. They were presented as in Figure 4.2.

Procedure. Subjects were seated in front of personal computers in groups of four or three. After receiving experimental instructions (see Appendices C and D), subjects answered the experimental questions. They were asked two practice choice questions to familiarize them with the experimental procedure. Subjects indicated their choice by clicking on the appropriate button. They could answer at their own pace. The experiment took 20 minutes on average.

Motivating subjects. Each subject received a flat fee of €3 for participation at the end of the experiment, plus a performance contingent payment (Eq. 4.1). The latter depended on the development of the prices for real estate in the next month. We told subjects that some of their choices concerned real data, and that the decision implemented at the end would concern the real data. There were in fact six choices with real data (in M^1 , M^2 , M^4 , M^5 , M^6 , and M^8), but subjects did not know which or how many these were. At the start of the experiment each subject was handed an envelope, which contained one of the memories with real data. The envelope was opened at the end of the experiment and the choice the subject had made in the question in the envelope was implemented. Implementation meant that we waited until next month's real estate appreciation had become known, after which the resulting outcome was transferred to the subjects' bank accounts. Given that the experimenters are professors at the same university where the students are, this procedure is trustworthy.

Table 4.1 shows that our questions were chained and an answer to one question might influence the next questions asked. However, the choice implemented for real was the one with the real values and this was independent of answers given by the subject. Hence subjects could understand that they could not benefit from strategic answering, and that it was in their interest to truthfully reply to all questions.

Consistency checks. The choices with real data were the same in M^1 and M^2 , in M^4 and M^5 , and in M^6 and M^8 (see Table 4.2). We used these three repeated choices to check the consistency of subjects' choices and to obtain an estimate of the error in their choices.

5.4.4 Predictions of CBDT

We next list three predictions of CBDT regarding comparisons of decision weights. In each prediction, part (a) is a prediction about the ratio of decision weights and the other parts are predictions about favorability, i.e. about the sign of the decision weights. We will, therefore, refer to Predictions 1a, 2a, and 3a as *ratio* predictions and to the other predictions as *sign predictions*.

The first prediction follows from a comparison between Eqs. 4.5. and 4.6, and tests linearity of utility. If utility is linear then the two ratios defined there should be equal. The other conditions of CBDT are less critical here because the memories M^1 and M^3 contain only c_0 and c_1 .

PREDICTION 1 OF CBDT (LINEAR UTILITY W.R.T. d_1): (a) $\frac{d_1^1}{d_0^2} = \frac{d_1^3}{d_0^1}$ (Eqs. 4.5 and 4.6). (b)

r_1^1 and r_1^3 imply the same sign (favorability) of d_1 . \square

The next two predictions test CBDT's assumption of separable cases (GSW Condition A5, GS Condition A2 p. 66, combination, and many related conditions). Prediction 2 follows from a comparison between Eqs. 4.5 and 4.7. If cases are separable then the common value of r_2 is irrelevant and the statistics defined in these equations should be the same. In memories M^1 and M^2 , r_2 is not specified, and in memories M^6 and M^7 it equals r_2 . In both cases, CBDT's separability implies that r_2 can be ignored, from which the predicted equality follows.

PREDICTION 2 OF CBDT (SEPARABILITY W.R.T. A COMMON r_2): (a) $\frac{d_1^1}{d_0^2} = \frac{d_1^6}{d_0^7}$ (Eqs. 4.5

and 4.7); (b) r_0^2 and r_0^7 imply the same sign of d_0 . (c) r_1^1 and r_1^6 imply the same sign of d_1 . \square

Prediction 3 follows from a comparison between Eqs. 4.8 and 4.9. Here the common value is r_1 . In memories M^4 and M^5 , r_1 is not specified. In memories M^7 and M^8 it is equal to r_1 . Now CBDT's separability implies that r_1 can be ignored, from which the predicted equality follows. According to separability of cases, the ratios defined in Eqs. 4.8 and 4.9 should be the same.

PREDICTION 3 OF CBDT (SEPARABILITY W.R.T. A COMMON r_1): (a) $\frac{d_2^5}{d_0^4} = \frac{d_2^8}{d_0^7}$ (Eqs. 4.8 and 4.9); (b) r_0^4 and r_0^7 imply the same sign of d_0 . (c) r_2^5 and r_2^8 imply the same sign of d_2 . \square

5.4.5 Analysis

If a decision weight is (close to) 0, then the subject never changes preference. Unfortunately, such choices may also arise due to a lack of attention on the subjects' part. In either case we cannot determine the sign of d_1 . In theoretical papers, such difficulties are usually avoided by assuming that all or at least several circumstances are nonneutral, often through various diversity axioms (for another condition, see GSW condition C5). We will treat zero values as missing.

Testing ratios is problematic if the denominator can be 0 or can change sign between subjects (Koerkamp et al. 2007). For example, the ratio $(-1)/(-2)$ is not the same as the ratio $1/2$. Fortunately, d_0 was negative for nearly all subjects, as expected, ($\geq 82\%$ for every measurement of d_0). We could, therefore, test the ratio predictions restricting them to the subjects for whom d_0 was negative. It is plausible that the few positive observations of d_0 were mostly due to error. The rate of positive weights d_0 was approximately equal to the observed inconsistency rates. Thus we could use the negative d_0 's to normalize the other decision weights, avoiding the complexities of analyzing ratios for which the denominator changes sign.

Because the signs of d_1 and d_2 were less consistent between subjects than those of d_0 , we did not use these for normalization purposes. That is, we did not compare the values of d_0 normalized by means of either d_1 or d_2 . We only compared the absolute strength of d_0 versus d_1 and d_2 by testing whether the absolute values $|d_j/d_0|$ were above or below 1 ($j = 1,2$).

All tests are two-sided. We used sign-tests to statistically test ratio predictions³¹. Within-subject (dependent samples) tests were used whenever possible. The rest of this paragraph explains our choice. Sign tests are conservative (have little power) but have the advantage of being applicable to the most general scales. Kolmogorov-Smirnov tests showed that we could not use t-tests. Wilcoxon signed-rank tests require comparability of differences of the scale between subjects. Given the novelty of the sign-dependent ratios of decision weights, little is known about their statistical properties, and the required comparability is questionable. Hence we chose to use the sign tests. The sign tests had enough power to detect differences in our data. When testing signs of similarity weights we used the usual binomial tests, within subjects whenever possible.

5.4.6 Results: Tests of CBDT

The consistency checks gave similar results for all three tests, averaging 23% inconsistencies for all choices, and 15% if zero decision weights were excluded. Such inconsistency rates are common in decision theory (Abdellaoui 2000; Camerer 1989; Harless & Camerer 1994; Hey & Orme 1994). The result is reassuring given that the stimuli for CBDT are more complex than those used in classical decision tasks. Subjects should not only consider the two acts to choose from but they should also consider varying memories containing multiple circumstances (Guerdjikova 2008 p. 115 l.-3).

Consistency of signs was generally confirmed. The null of random signs could be rejected in favor of consistent signs ($p < 0.01$) in all but one case (d_1^6 in Prediction 2c). In Predictions 1b, 2b, 3b, and 3c the null of no sign changes could be accepted (always $p > 0.3$). This null could only be rejected in Prediction 2c ($d_1^1 > d_1^6$; $p = 0.01$).

For the ratio predictions, Predictions 1a and 3a were accepted ($p > 0.5$), but Prediction 2a was rejected ($p = 0.007$). That is, linear utility was accepted, and separability of cases was accepted when (c_1, r_1) was added to memory, but not when (c_2, r_2) was added to memory.

³¹ By the sign test we mean the well know distribution-free statistical test, it has nothing to do with

5.4.7 Results: Explorations regarding real estate investments

Our expectations about the signs and the strengths of the decision weights were confirmed by the data. We anticipated that c_2 would be maximally favorable for a_1 , that c_0 would be maximally unfavorable for a_1 , and that c_1 would be in between. We did not have a clear expectation about the sign of c_1 . We indeed observed $d_0 < 0$, $d_2 > 0$, and $d_2 > d_1$ ($\frac{d_2^5}{d_0^4} > \frac{d_1^1}{d_0^2}$; $p = 0.02$; $\frac{d_2^8}{d_0^7} > \frac{d_1^6}{d_0^0}$ ($p = 0.05$). As explained, we could not directly compare d_0 with d_1 and d_2 , but we could compare the strength of d_0 with the strengths of the other weights. We found that $|d_1/d_0| < 1$ ($|\frac{d_1^1}{d_0^2}| < 1, |\frac{d_1^3}{d_0^1}| < 1$, and $|\frac{d_1^6}{d_0^0}| < 1$, all with $p < 0.001$), but $|d_2/d_0| = 1$ ($p = 1$, both for $|\frac{d_2^5}{d_0^4}|$ and $|\frac{d_2^8}{d_0^0}|$). Thus c_0 and c_2 had similarly strong effects (although in opposite directions) and both had more effect than c_1 did. The sign of d_1 was positive (although not significant in M^6). Apparently geographical vicinity (Limburg borders North-Brabant) was more important than the difference in the type of dwelling (house versus apartment).

5.4.8 Discussion of Experiment 1

The predictions of CBDT were accepted with one exception: separability of cases for d_1 if (c_2, r_2) is added to memory (M^1 versus M^6), in Prediction 2. This violation means that the effect of prices of a house in Limburg is affected by information about the result of an apartment in Utrecht. This violation of separability is not surprising. In M^1 the only other circumstance in memory is c_0 , which is similar to a_0 and favors it. It then is natural that c_1 is taken to favor a_1 . In M^6 , c_2 is also present in memory. It concerns the same type of dwelling (apartment) as a_1 making it more questionable whether c_1 , which concerns a different type of dwelling (a house), should favor a_1 . The inclusion of c_2 weakens the support of c_1 for a_1 . This concerns a plausible interaction between circumstances in memory, which violates the separability of CBDT.

Despite the one violation of CBDT, its measurements give useful insights into real estate investment decisions. The preference for a_1 over a_0 is primarily affected by negative results for c_0 , a bit less (although not significantly so) by positive results for c_2 , and the least strongly by positive results for c_1 . These findings are plausible. For c_1 we learned from the experiment that subjects relate it more to a_1 than to a_0 . Using the terminology of GS (p. 78), the attribute of geographic similarity plays a bigger role than the attribute of type-of-dwelling similarity does.

5.5 Experiment 2 to measure similarity weights and to test CBDT

In Experiment 1, circumstance c_1 was complex to process for subjects because type of dwelling and geography had opposite effects. This may have contributed to the observed violation of separability and to the inconsistencies in choice. Hence, Experiment 2 used cases that were easier to process. Favorability was always unquestionable and geography and type of dwelling always had the same effect. In addition, there were fewer subjects per session. Both changes reduced noise and increased statistical power. Unfortunately, because this experiment had to be done in the same month as Experiment 1, we could only obtain a limited number of subjects. In what follows, we will focus on the differences between the two experiments.

5.5.1 Stimuli to measure similarity weights

We used the same acts a_0 and a_1 to choose from and the same payment scheme. The circumstances in memory were:

c_0 (house in North-Holland);

c_3 (apartment in Limburg);

c_4 (apartment in Gelderland).

The decision weights d_0 , d_3 , and d_4 are as in Eq. 2.1. Circumstance c_3 more clearly support a_1 against a_0 than c_1 did in Experiment 1: c_3 concerns the same type of dwelling as a_1 did, whereas c_1 concerned a different type. It is also easier for subjects to compare c_3 and c_4 in the three-circumstance memories (M^{14} , M^{15} , M^{16}), because they only differ geographically and not regarding type of dwelling. The real annual price increases of c_0 , c_3 , and c_4 over the past 3 years were 15.9%, 4.1%, and 6.0%.

Table 5.1 depicts the design of the second experiment, which is of the same for as in Experiment 1. As in Experiment 1, the outcomes in bold were varied to produce indifference and the outcomes not in bold were set equal to their real values.

TABLE 5.1

	M^9	M^{10}	M^{11}	M^{12}	M^{13}	M^{14}	M^{15}	M^{16}
c_0	r_0	r_0^{10}	r_3^9+5	r_0^{12}	r_0	r_0	r_0^{15}	r_0
c_3	r_3^9	r_3	r_3^{11}			r_3^{14}	r_3	r_3
c_4				r_4	r_4^{13}	r_4	r_4	r_4^{16}

$$M^9 = \{(c_0, r_0), (c_3, r_3^9)\} \text{ and } M^{16} = \{(c_0, r_0), (c_3, r_3), (c_4, r_4^{16})\}.$$

A difference with Experiment 1 was that in M^{11} we added 5 to r_3^9 rather than to r_0^{10} , to test linear utility for other differences in outcomes. By Theorem 2.1 we have the following results, using the same shorthand notation as in Experiment 1.

$$\frac{d_3}{-d_0} = \frac{r_0 - r_0^{10}}{r_3^9 - r_3} =: \frac{d_3^9}{d_0^{10}} (M^9 \text{ \& } M^{10}). \quad (5.1)$$

$$\frac{d_3}{-d_0} = \frac{r_0 - (r_3^9 + 5)}{r_3^9 - r_3^{11}} =: \frac{d_3^{11}}{d_0^9} (M^9 \text{ \& } M^{11}). \quad (5.2)$$

$$\frac{d_3}{-d_0} = \frac{r_0 - r_0^{15}}{r_3^{14} - r_3} =: \frac{d_3^{14}}{d_0^{15}} (M^{14} \text{ \& } M^{15}). \quad (5.3)$$

$$\frac{d_4}{-d_0} = \frac{r_0 - r_0^{12}}{r_4^{13} - r_4} =: \frac{d_4^{13}}{d_0^{12}} (M^{12} \text{ \& } M^{13}). \quad (5.4)$$

$$\frac{d_4}{-d_0} = \frac{r_0 - r_0^{15}}{r_4^{16} - r_4} =: \frac{d_4^{16}}{d_0^{15}} (M^{15} \text{ \& } M^{16}). \quad (5.5)$$

5.5.2 Sample and procedure

Subjects. $N = 23$ (12 female) undergraduate students from Erasmus University signed up for the experiment. The subjects had various academic backgrounds.

Procedure. The experiment was administered through individual interviews or in sessions involving 2 students. Thus, there were fewer subjects per session than there were in Experiment 1.

Consistency checks. The real choices are the same in M^9 and M^{10} , in M^{12} and M^{13} , and in M^{14} and M^{16} . We used these pairs to test consistency.

5.5.3 Predictions of CBDT

We could again derive three predictions from CBDT, one testing linearity of utility and the other two testing separability of cases.

PREDICTION 4 OF CBDT (LINEAR UTILITY W.R.T. d_3): (a) $\frac{d_3^9}{d_0^{10}} = \frac{d_3^{11}}{d_0^9}$ (Eqs. 5.1 and 5.2) (b) r_3^9 and r_3^{11} imply the same sign of d_3 . \square

PREDICTION 5 OF CBDT (SEPARABILITY W.R.T. A COMMON r_4): (a) $\frac{d_3^9}{d_0^{10}} = \frac{d_3^{14}}{d_0^{15}}$ (Eqs. 5.1 and 5.3). (b) r_0^{10} and r_0^{15} imply the same sign of d_0 . (c) r_3^9 and r_3^{14} imply the same sign of d_3 . \square

PREDICTION 6 OF CBDT (SEPARABILITY W.R.T. A COMMON d_3): (a) $\frac{d_4^{13}}{d_0^{12}} = \frac{d_4^{16}}{d_0^{15}}$ (Eqs. 5.4 and 5.5). (b) r_0^{12} and r_0^{15} imply the same sign of d_0 . (c) r_4^{13} and r_4^{16} imply the same sign of d_4 . \square

5.5.4 Results: Tests of CBDT

The consistency checks gave similar results on the three tests, averaging 7% inconsistencies overall. This is considerably better than in Experiment 1.

All predictions of CBDT were accepted. The null of random signs could always be rejected in favor of consistent signs. The sign and ratio predictions of CBDT were always accepted, with always $p > 0.15$.

5.5.5 Results: Explorations regarding real estate investments

All results in Experiment 2 are plausible and are as anticipated. We have $d_0 < 0$, $d_3 > 0$, and $d_4 > 0$ (always $p < 0.01$). We find no difference of strength between d_3 and d_4 , with $\frac{d_4^{13}}{d_0^{12}} = \frac{d_3^9}{d_0^{10}}$ and $\frac{d_4^{16}}{d_0^{15}} = \frac{d_3^{14}}{d_0^{13}}$ (always $p > 0.6$). We found $|d_3/d_0| < 1$ and $|d_4/d_0| < 1$

$(|\frac{d_4^{14}}{d_0^{13}}| < 1, p = 0.02; |\frac{d_3^{16}}{d_0^{15}}| < 1, p = 0.02)$, indicating that c_0 provided more support for

a_0 than both c_3 and c_4 did for a_1 , although the inequalities were not significant for $(|\frac{d_3^9}{d_0^{10}}|$
 $)$ and $(|\frac{d_4^{13}}{d_0^{12}}|)$.

5.5.6 Discussion of Experiment 2

All predictions of CBDT were accepted in the second experiment. One might worry that this could be due to lack of power, because the number of subjects was less than half the number in Experiment 1. However, in this experiment we made a special effort to have clear cases in memory with clear predictions, and to reduce noise, reducing inconsistency by a factor 3. All findings are plausible. The signs of the decision weights were significantly different than 0, and the strengths (irrespective of direction) of c_3 and c_4 were smaller than c_0 's strength, with no other significant differences. The equality of d_3 and d_4 suggests that Limburg and Gelderland are to the same extent more similar to Brabant than to South-Holland. The preference of a_1 over a_0 was most affected by negative results for c_0 and then, and in equal measure, by positive results for c_3 and c_4 .

5.5.7 A Comparison Between Experiment 2 and Experiment 1

We compared all ratios d_j/d_0 of Experiment 2 with those of Experiment 1 and found the following significant differences.

$$d_3 > d_1 \left(\frac{d_3^9}{d_0^{10}} > \frac{d_1^1}{d_0^2}, p = 0.02; \frac{d_3^{14}}{d_0^{15}} > \frac{d_1^6}{d_0^7}, p = 0.005 \right).$$

$$d_4 > d_1 \left(\frac{d_4^{16}}{d_0^{15}} > \frac{d_1^6}{d_0^7}, p = 0.005 \right).$$

These differences are plausible. In c_3 and c_4 the type of dwelling was the same as in a_1 while it was different in c_1 , and geographical differences with a_1 were identical or comparable for c_1 , c_3 , and c_4 . All other ratios d_j/d_0 did not differ significantly between the two experiments. Again, this was according to expectation because the differences in type of dwelling and geography were always the same.

Combining the two experiments, we find that the preference for a_1 over a_0 was most affected by negative results for c_0 , then by positive results for c_2 , then by positive results for c_3 and c_4 , and, finally, to the least extent by positive results for c_1 .

5.6 General discussion

Most assumptions of CBDT were corroborated by our tests. The only exception was a violation of separability of different cases in memory: The informational value of a house in Limburg (c_1) was affected by that of an apartment in Utrecht (c_2). As pointed out by Gilboa & Schmeidler (1995, p. 631; 1997a p. 52), such a violation is similar to the violations of separability over disjoint events (the sure-thing principle, or independence) found for expected utility, and is equally unsurprising in retrospect. Further violations of separability are discussed by GS (p. 74).

Although our violation of separability was found under a change of memory size, it reflects an interaction between cases that will also generate violations when memories are of the same size. This could, for instance, be tested by adding neutral cases to the smallest memories. Eichberger & Guerdjikova's (2010, 2011) developed restrictions of separability to memories of equal size, and imposed mixture versions of

independence for such cases (generalizing the concatenation axiom of Billot et al. 2005). We conjecture that the interaction of cases that we found can also lead to violations of those weakened axioms of separability, but leave the actual investigation of this claim to future studies.

Gayer, Gilboa, & Lieberman (2007) found that case-based reasoning played a bigger role in the rental market than in the more speculative sales market. This suggests that we have put CBDT to a hard test, investigating it in a domain, the market for real-estate, where its effects were previously found to be rather weak. We nevertheless observed clear support for the predictions of CBDT. Our finding does not contradict Gayer Gilboa, & Lieberman (2007) because our subjects were price takers, unlike those involved in the sales market, and the gains for our subjects depended on the market price. They were not involved in price negotiations.

We, finally, discuss the implications of our results for some versions of CBDT that are alternative to the one used in our chapter. Special cases of the act similarity version in Eq. 1.1 arise if similarity s depends only on the problems p and q (GS p. 35; Gilboa & Schmeidler 1995 p. 610), if s is 0 whenever $b \neq a$ (GS p. 38; Gilboa & Schmeidler 1995 p. 610), and if each q appears at most once in M (GS pp. 37-38; Gilboa & Schmeidler 1995). Because these are special cases of Eq. 1.1 our measurements are also valid under them. Further generalizations occur when circumstances in memory are not decomposed into problems and acts (to which our analysis applies with no modification), and when similarity can also depend on the result (GS p. 52). The latter dependency is too general for its parameters to be measurable unless we can use in the repetitions approach where each circumstance (q,b) can occur any finite number of times in memory (GS Ch. 3). The data set in our experiment is not rich enough for the repetitions approach, and we do not consider it in this chapter.

A final alternative version of CBDT arises if the similarity weights are normalized, leading to an average rather than a (weighted) sum of utility. This approach is appropriate if we decide infinitely often, using the present choice merely to find the long-term highest average. Our chapter and experiment consider one-off choices.

Then maximizing the sum in Eq. 1.1, and not the average, is appropriate (GS pp. 74, 158 ff.; Pape & Kurtz 2012).

5.7. Summary and Conclusion

This chapter has introduced a parameter-free method to measure similarity weights, the main new components of case-based decision theory (GS p. 35). This method directly shows the relation between the weights and decisions, without imposing any restrictions on either. We assumed linear utility, which is reasonable for the moderate amounts used in our experiment. An extensions to nonlinear utility is in the appendix.

Our measurement method works as follows. If a preference for a_1 over a_0 can be turned into an indifference by increasing the outcome (appreciation) under c_i by ϵx , and also by increasing the outcome under c_j by ϵy , then the ratio d_i/d_j of decision weights is y/x . It is a case-based analog of de Finetti's betting odds for nonparametrically measuring subjective probabilities.

Decision weights are differences of similarity weights and only their ratios and signs can be observed. CBDT generally requires information on several unknowns (the neutral utility level and the cases in memory prior to the experiment), which may be hard to obtain. An advantage of our method is that we do not need this information because these unknowns dropped from our equations.

We developed a case-based analog of the random incentive system. Thus we could manipulate memories in an incentive-compatible manner without requiring subjects to forget information once given. We resolved the statistical complication of testing ratios with changing signs by separating sign- and ratio predictions. An experiment showed that our method is implementable. Case-based decision theory was generally supported by our data, although generalizations of separability of cases in memory may be desirable.

We found the predictions of CBDT reasonably well confirmed, although generalizations of separability of cases in memory are desirable. Case-based decision theory is a viable alternative to classical revealed preference. It agrees with intuitive similarity judgments (based on Dutch conditions in our experiment), and gives new

insights where intuitive judgments are inconclusive. Thus it gave new insights into (real estate investment) decisions. Our experiment is easy to implement and transparently shows the empirical meaning of similarity weights. We hope that this chapter will encourage further empirical investigations into CBDT.

Appendix A: Utility measurement under CBDT

GSW proposed two methods for measuring utility. One (their §3) adopts the repetitions approach, where circumstances can be repeated and thus weighted differently.³² Since we do not have such data available in this study, we do not discuss it further. For the second method, we consider variations of outcomes of two fixed circumstances in memory, an assumption made in the rest of this section. We thus consider

$$M = \{(q_0, c_0, r_0), (q_1, c_1, r_1)\} \cup M_p, \text{ denoted } (r_0, r_1), \quad (\text{A.1})$$

for various (r_0, r_1) . The notation (r_0, r_1) , called a context in GSW and used only in this appendix, can work because all other variables are kept fixed here. Under some nondegeneracy assumptions, indifferences

$$a_0 \sim_{\{\alpha_0, \sigma_1\}} a_1, a_0 \sim_{\{\beta_0, \tau_1\}} a_1, a_0 \sim_{\{\gamma_0, \sigma_1\}} a_1, a_0 \sim_{\{\delta_0, \tau_1\}} a_1, \quad (\text{A.2})$$

imply

$$u(\alpha_0) - u(\beta_0) = u(\gamma_0) - u(\delta_0). \quad (\text{A.3})$$

This can be derived from substitution of Eq. 1.1, as was demonstrated by GSW (Eq. 6). The indifferences in Eq. A.2 are such that all unknowns from the equations, such as decision/similarity weights and the terms referring to M_p , drop. Under usual richness assumptions, equalities of utility differences suffice to measure utility u up to level and unit. To completely measure utility, we should also determine its level (i.e., where utility is 0), which can be inferred by verifying the neutrality condition.

³² Another, distribution-based, way of weighting cases unequally is proposed in similarity-based forecasting (Lovallo, Clarke, & Camerer (2012)).

However, this is not needed for the measurement of similarity weights, the new parameters of CBDT.

Appendix B: Proof

PROOF OF THEOREM 2.1. The first indifference in Eq. 2.3 implies

$$\begin{aligned} & s((p,a_1),(q_0,c_0))u(r_0^2) + s((p,a_1),(q_1,c_1))u(r_1) + \sum_{(q,b,r) \in M_p} s((p,a_1),(q,b))u(r) = \\ & s((p,a_0),(q_0,c_0))u(r_0^2) + s((p,a_0),(q_1,c_1))u(r_1) + \sum_{(q,b,r) \in M_p} s((p,a_0),(q,b))u(r). \end{aligned}$$

Substituting Eq. 2.1 and writing $K = \sum_{(q,b,r) \in M_p} s((p,a_1),(q,b))u(r) -$

$\sum_{(q,b,r) \in M_p} s((p,a_0),(q,b))u(r)$, we get

$$d_0u(r_0^2) + d_1u(r_1) + K = 0. \quad (\text{B.1})$$

The second indifference in Eq. 2.3 similarly implies

$$d_0u(r_0) + d_1u(r_1^1) + K = 0. \quad (\text{B.2})$$

Subtracting Eq. B.2 from Eq. B.1 gives

$$d_0(u(r_0^2) - u(r_0)) - d_1(u(r_1^1) - u(r_1)) = 0. \quad (\text{B.3})$$

This implies Eq. 2.4, where both numerators and denominators are nonzero because of Eq. 2.3. The results on the signs of d_0 and d_1 follow from Eq. 1.1. \square

Appendix C: Instructions for subjects

You are participating in an experiment on decision making. During the experiment, we will provide you with information about the development of prices of some real estate types in different regions of the Netherlands in the last three years. Based on the information presented in each of the questions, you have to choose between two types of property to gamble on: either a single-family house (in Dutch: eengezinswoning) in

the province of Zuid-Holland or an apartment (in Dutch: appartement) in the province of Noord Brabant.

You have just chosen a computer. Next to this computer lies an envelope. This envelope will be opened by the experimenter at the end of the experiment. It contains true information about the development of real estate prices in the Netherlands between 2005 and 2008 (from the Dutch Cadaster Index). It also offers you a choice between two types of property to gamble on. You will face this particular choice at some point during the experiment. You will receive a payment that is based on your choice during the experiment and on the actual movements of real estate prices (houses or apartments) in the month of the experiment (March 2009). It should be emphasized that neither you nor the experimenters know at this moment how the housing market will evolve in this month and, therefore, what your payment will be in a month.

Because we are interested in your choices in many situations, we will ask you several questions. In each question, you receive a piece of information and you will be asked to make a what-would-you-gamble-on-if choice, based on this piece of information. Often this information is hypothetical and made up by us. But, as mentioned before, one of the questions asked during the experiment is the one contained in the envelope and is based on real data.

For each question, only the piece of information provided there is relevant, and you best decide assuming that this is the only piece of information you got. After all, if this question turns out to be the one contained in your envelope, then all the pieces of information provided in the other questions are not true and are irrelevant for the payment you will receive. That is, best for you to do at each question is to forget all information provided in previous questions, and to focus only on the information provided at that question now faced.

Your payment is determined as follows. In addition to a show-up fee of €3, you will get another €3 + 10 times the monthly appreciation rate (in percentage) of the property you chose in the question contained in the envelope. In case the property value will decrease, we will subtract it from these additional €3, but not from the original show-up fee. You will always receive at least the show-up fee. Please bear in

mind that despite the slowdown of the economy, real estate prices in the Netherlands still went up during the last months.

EXAMPLE:

For example, assume that you choose an apartment in Zuid-Holland in the real question. If the prices of apartments in Zuid-Holland go up by 0.6% in March, you receive $\text{€}3 + 0.6\% \times 10 = \text{€}3 + \text{€}6 = \text{€}9$, plus the original show-up fee of $\text{€}3$. If the prices go down by 0.2%, you get $\text{€}3 - 0.2\% \times 10 = \text{€}3 - \text{€}2 = \text{€}1$ for your investment, plus the original show-up fee of $\text{€}3$. If the prices go down by 0.3% or more, you get $\text{€}0$ for your investment but you still keep the show-up fee of $\text{€}3$.

Once the Cadaster makes the information about the real estate prices in March public (in April 2009), we will inform you by e-mail and either deposit the money in your bank account, or you can collect it in the office of the experimenter, L3-121.

References

- Abdellaoui, M. (2000). Parameter-Free Elicitation of Utility and Probability Weighting Functions. *Management Science* 46, 1497–1512.
- Aha, D. W., Marling, C. & Watson, I. D. (2005, eds.). The Knowledge Engineering Review, Special Edition on Case-Based Reasoning. 20, *Cambridge University Press*, Cambridge UK.
- Anscombe, F. J. & Aumann, R. J. (1963). A Definition of Subjective Probability. *Annals of Mathematical Statistics* 34, 199–205.
- Arrow, K. J. (1951). Social Choice and Individual Values. Wiley, New York. (9th edn. 1972, *Yale University Press*, New Haven.)
- Aumann, R. J. (1971, January 8). Letter from Robert Aumann to Leonard Savage. Published as Appendix A to Ch. 2 of Jacques H. Drèze (1987). Essays on Economic Decision under Uncertainty. *Cambridge University Press*, Cambridge.

- Billot, A., Gilboa, I., Samet, D. & Schmeidler, D. (2005). Probabilities as Similarity-Weighted Frequencies. *Econometrica* 73, 1125–1136.
- Blonski, M. (1999). Social Learning with Case-Based Decisions. *Journal of Economic Behavior and Organization* 38, 59–77.
- Camerer, C. F. (1989). An Experimental Test of Several Generalized Utility Theories. *Journal of Risk and Uncertainty* 2, 61–104.
- Charness, G. & Levin, D. (2005). When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect. *American Economic Review* 95, 1300–1309.
- de Finetti, B. (1931). Sul Significato Soggettivo della Probabilità. *Fundamenta Mathematicae* 17, 298–329. Translated into English as “On the Subjective Meaning of Probability,” in Paola Monari & Daniela Cocchi (1993, eds.) “*Probabilità e Induzione*,” 291–321, Clueb, Bologna.
- Dubois, D., Godo, L., Prade, H. & Zapico, A. (1999). On the Possibilistic Decision Model: From Decision under Uncertainty to Case-Based Decision. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 7, 631–670.
- Eichberger, J. & Guerdjikova, A. (2010). Case-Based Belief Formation under Ambiguity. *Mathematical Social Sciences* 60, 161–177.
- Eichberger, J. & Guerdjikova, A. (2011). Ambiguity, Data and Preferences for Information— A Case-Based Approach. Mimeo.
- Gayer, G. (2010). Perception of Probabilities in Situations of Risk; A Case Based Approach. *Games and Economic Behavior* 68, 130–143.
- Gayer, G., Gilboa, I. & Lieberman, O. (2007). Rule-Based and Case-Based Reasoning in Housing Prices. *B.E. Journal of Theoretical Economics* 7, Iss. 1 Article 10.
- Gilboa, I. (2009). Theory of Decision under Uncertainty. Econometric Society Monograph Series, *Cambridge University Press*, Cambridge.

- Gilboa, I., Lieberman, O. & Schmeidler, D. (2010). On the Definition of Objective Probabilities by Empirical Similarity. *Synthese* 172, 79–95.
- Gilboa, I., Lieberman, O. & Schmeidler, D. (2011). A Similarity-Based Approach to Prediction. *Journal of Econometrics* 162, 124–131.
- Gilboa, I. & Schmeidler, D. (1995). Case-Based Decision Theory. *Quarterly Journal of Economics* 110, 605–639.
- Gilboa, I. & Schmeidler, D. (1996). Case-Based Optimization. *Games and Economic Behavior* 15, 1–26.
- Gilboa, I. & Schmeidler, D. (1997a). Act-Similarity in Case-Based Decision Theory. *Economic Theory* 9, 47–61.
- Gilboa, I. & Schmeidler, D. (1997b). Cumulative Utility Consumer Theory. *International Economic Review* 38, 737–761.
- Gilboa, I. & Schmeidler, D. (2001). A Theory of Case-Based Decisions. *Cambridge University Press*, Cambridge.
- Gilboa, I., Schmeidler, D. & Wakker, P. P. (2002). Utility in Case-Based Decision Theory. *Journal of Economic Theory* 105, 483–502.
- Golosnoy, V. & Okhrin, Y. (2008). General Uncertainty in Portfolio Selection: A Case-Based Decision Approach. *Journal of Economic Behavior and Organization* 67, 718–734.
- Greco, Salvatore, Benedetto Matarazzo & Slowinski, R. (2008). Case-Based Reasoning Using Gradual Rules Induced from Dominance-Based Rough Approximations. *Springer*, Berlin.
- Greenspan, A. (2004). Innovations and Issues in Monetary Policy: The Last Fifteen Years. *American Economic Review, Papers and Proceedings* 94, 33–40.
- Guerdjikova, A. (2008). Case-Based Learning with Different Similarity Functions. *Games and Economic Behavior* 63, 107–132.

- Harless, D. W. & Camerer, C. F. (1994). The Predictive Utility of Generalized Expected Utility Theories. *Econometrica* 62, 1251–1289.
- Hertwig, Ralf, Barron, G., Weber, E. U. & Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science* 15, 534–539.
- Hey, J. D. & Orme, C. (1994). Investigating Generalizations of Expected Utility Theory Using Experimental Data. *Econometrica* 62, 1291–1326.
- Holt, C. A. (1986). Preference Reversals and the Independence Axiom. *American Economic Review* 76, 508–513.
- Hüllermeier, E. (2007). Case-based Approximate Reasoning. *Springer*, Berlin.
- Jahnke, H., Chwolka, A. & Simons, D. (2005). Coordinating Service-Sensitive Demand and Capacity by Adaptive Decision Making: An Application of Case-Based Decision Theory. *Decision Sciences* 36, 1–32.
- Karni, E. & Vierø, M.-L. (2011). Reverse Bayesianism: A Choice-Based Theory of Growing Awareness. Mimeo.
- Keeney, R. L. & Raiffa, H. (1976). *Decisions with Multiple Objectives*. Wiley, New York (2nd edn. 1993, *Cambridge University Press*, Cambridge).
- Grosskopf, B., Sarin, R. & Watson, E. (2008). An Experiment on Case-Based Decision Making. Mimeo.
- Groot K., Bas, M. G., Hunink, M., Stijnen, T., Hammitt, J. K., Kuntz, K. M., Weinstein, M. C. (2007). Limitations of Acceptability Curves for Presenting Uncertainty in Cost-Effectiveness Analysis. *Medical Decision Making* 27, 101–111.
- Lovallo, D., Clarke, C. & Camerer, C. (2012). Robust Analogizing and the Outside View: Two Empirical Tests of Case-Based Decision Making. *Strategic Management Journal* 33, 496–512.

- Luce, R. D. (2000). *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches*. Lawrence Erlbaum Publishers, London.
- Manski, C. F. (2011). Actualist Rationality. *Theory and Decision* 71, 297–324.
- Pape, A. D & Kurtz, K. J. (2012). Evaluating Case-Based Decision Theory: Predicting Empirical Patterns of Human Classification Learning. Mimeo.
- Peski, M. (2011). Prior Symmetry, Similarity-Based Reasoning, and Endogenous Categorization. *Journal of Economic Theory* 146, 111–140.
- Ramsey, F. P. (1931). Truth and Probability. In Richard B. Braithwaite (ed.), *The Foundations of Mathematics and other Logical Essays*, 156–198, Routledge and Kegan Paul, London.
Reprinted in Henry E. Kyburg Jr. & Howard E. Smokler (1964, eds.) *Studies in Subjective Probability*, 61–92, Wiley, New York. (2nd edn. 1980, Krieger, New York.)
- Riesbeck, C. K. & Schank, R. C. (1989). *Inside Case-Based Reasoning*. Lawrence Erlbaum, Hillsdale, NJ.
- Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York. (2nd edn. 1972, Dover Publications, New York.)
- Starmer, C. & Sugden, R. (1991). Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation. *American Economic Review* 81, 971–978.
- Stewart, N., Chater, N. & Brown, G. D. A. (2006). Decision by Sampling. *Cognitive Psychology* 53, 1–26.
- Wakker, P. P. (2010). *Prospect Theory: for Risk and Ambiguity*. Cambridge University Press, Cambridge, UK.

Chapter 6

Conclusion

This thesis tests several nonstandard decision theories in health and real estate domains. In general, once deviations from the standard expected-utility model are accounted for, the investigated conditions pass these tests on an aggregate level. Both the additive model for health sequences and similarity-based models for housing market decisions provide very good approximations of human decision making. For individual decision making, there are more deviations.

Tests of the QALY model and a generalization thereof do not require additional confounding assumptions because they reckon with violations of constant discounting and expected utility. At the aggregate level I observed support for the QALY model as its critical condition, generalized marginality, could not be rejected. At the individual level there was less support for the QALY model: a sizeable proportion of the subjects violated generalized marginality. The observed deviations were too large to be caused by elicitation and preference imprecisions alone.

I also evaluated utility independence empirically. It is a less restrictive preference condition than generalized marginality, but still implies a tractable model. Utility independence was supported well at the aggregate level. At the individual level, I found more support for utility independence than for generalized marginality. For a substantial proportion of our subjects the observed violations of utility independence can be attributed to the elicitation procedure and preference imprecision.

The results just mentioned suggest that QALYs cannot be applied in individual medical decision-making without some additional tests of the decision maker's preference structure. Laying out this preference structure is an interesting topic for further research.

Even though we found violations of the QALY model at the individual level, not all is lost. There was more support for utility independence not only at the aggregate, but

also at the individual level. Utility independence still implies a tractable model that can be applied in practice.

Most assumptions of Case-Based Decision Theory (CBDT) were corroborated by our tests. The only exception was separability of different cases in memory, which was violated on one occasion: the informational value of a house in one of the Dutch provinces was affected by that of an apartment in the other. Such a violation is similar to the violations of separability over disjoint events (the sure-thing principle, or independence) found for expected utility.

The research outlined in this thesis raises some questions – both philosophical and practical – about feasibility and appropriateness of laboratory experiments in testing axiomatic foundations of human decision making. The results show that important insights can be obtained by letting subjects answer seemingly simple questions.