In search of a preferred preference elicitation method: A test of the internal consistency of choice and matching tasks

Attema, Arthur and Brouwer, Werner

Erasmus University Rotterdam, Erasmus University Rotterdam

20 January 2012

In search of a preferred preference elicitation method

A test of the internal consistency of choice and matching tasks

Arthur E. Attema[a] and Werner B.F. Brouwer[b]

[a] (Corresponding author) iBMG/iMTA, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands. E-mail: attema@bmg.eur.nl, --31-10.408.91.29 (O); --31-10.408.90.81 (F)
[b] iBMG/iMTA, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands. E-mail: brouwer@bmg.eur.nl.

January, 2012

ABSTRACT. The numerous reports on preference reversals in preference elicitations pose a great challenge to empirical economics. Many studies have found that different procedures may generate substantially different preferences. However, little is known about whether one procedure is more susceptible to preference reversals than another. Therefore, taking the preference reversals as a robust behavioral pattern, guidelines are called for to provide directions regarding a preferred preference elicitation task. This paper puts forward a new test of the internal consistency of choice and matching tasks, based on "internal preference reversals". We replicate the preference reversal phenomenon and find a significant higher consistency within choice tasks than within matching tasks.

Key Words: preference reversal, internal consistency, scale compatibility, loss aversion, time tradeoff method

JEL CLASSIFICATION: B40, C91, I10

1. Introduction

A major problem for classical decision theory is the frequent finding of *preference reversals*, i.e., the phenomenon that the relative evaluation of two or more items by an individual depends systematically on the elicitation method used (Cubitt et al., 2004). Starting with Lindman (1971) and Lichtenstein and Slovic (1971), preference reversals have been extensively investigated in the context of lotteries (Tversky and Thaler, 1990). The early studies on preference reversals compared a choice between two gambles and the selling prices of these gambles. If a specific gamble is preferred to another, economic theory predicts that this gamble should also be sold at a higher price than the other. If someone prefers prospect A over B, one would expect that person to be willing to pay more for A too. However, it turned out that, when offering one gamble with a high probability of winning a modest amount of money (a *P-bet*) and one gamble with a low probability of winning a relatively large amount of money (a *$-bet*), many individuals chose the first option but at the same time stated a higher price for the second one. Preference reversals were found in several other tasks as well, and in a variety of different environments (Seidl, 2002).

Because of the robust nature of this phenomenon, alternative theories were developed to explain it (Fishburn, 1984; Fishburn, 1985; Holt, 1986; Karni and Safra, 1987; Loomes and Sugden, 1982; Loomes and Sugden, 1983; Tversky et al., 1988). Tversky et al. (1990) empirically tested these theories in a carefully designed experiment, and found convincing evidence in favor of a theory that drops the procedural invariance assumption (Tversky et al., 1988). They claimed that the violation of procedural invariance to a large extent explained the preference reversal phenomenon. The main causes of violation of procedural invariance are

scale compatibility and loss aversion.[1] *Scale compatibility* means that people assign greater weight to attributes represented in units similar to those of the response variable, which can generate a large distorting influence (Borcherding et al., 1991; Delquié, 1993; Huber et al., 2002; Tversky et al., 1988). *Loss aversion* is the phenomenon that individuals handle gains and losses as seen from a reference point differently, with losses looming larger than gains (Tversky and Kahneman, 1991; Tversky and Kahneman, 1992).

In this paper we focus on the kind of preference reversal that is caused by differences in matching and choice tasks, because these are the two most common elicitation tasks. In a choice task, an individual has to choose one option, possibly consisting of multiple attributes, out of a set of options. In matching, one attribute in one of the options is left blank and the subject is asked to give a value for this attribute that makes him indifferent between the options. Instead of just testing for the existence of preference reversals, however, we pursue another line of research.

Since the literature on preference reversals is large, we can be confident that the task used in assessing preferences exerts a substantial influence on the outcomes (Maafi, 2011). But, even if we are certain about the existence and the causes of preference reversals, what does this imply for the practical use of different assessment procedures? Is there one task that elicits preferences more consistently than others? Indeed, while one may prefer revealed over stated preference, the latter is commonly used in situations where revealed preference data cannot be ethically or meaningfully obtained, such as in the context of health. Hence, the question of a preferred elicitation method remains pivotal.

The differences between choice and matching elicitation tasks have been investigated before (Bostic et al., 1990; Huber et al., 2002; Loomes, 1988; Stalmeier et al., 1997). The major conclusion from these studies was that elicitation of indifferences by means of choices

[1] Recently, however, (Butler and Loomes, 2007) and (Schmidt and Hey, 2004) casted doubt on the validity of the scale compatibility hypothesis and suggested that preference reversals are partly caused by a higher error

yields better results than indifferences obtained by matching, in the sense that series of choices generated fewer inconsistencies than matching questions (Bostic et al., 1990), although Loomes (1988) and Loomes and Sugden (1989) found preference reversals for both tasks in a monetary lottery setting. Arrow et al. (1993) recommended the use of choices instead of open-ended question in contingent valuation studies, because they consider the open-ended questions not realistic and sensitive to strategic responses.

This paper seeks to extend the above research by performing a specific test of the internal consistency of choice and matching tasks. We test this in a health valuation setting, where these procedures are frequently used (also to inform actual decisions). In particular, we use a time tradeoff (TTO) valuation, which is a popular method to elicit preferences for health states (Dolan et al., 1996a; Dolan et al., 1996b; Lamers et al., 2006) and has been used to derive value sets in a number of countries, including Denmark, Germany, Japan, the Netherlands, Spain and the UK (http://www.euroqol.org). In short, the method asks an individual to trade off life years in order to improve health status. However, there exist different, strategically equivalent, ways to elicit preferences with this method. This allows us to test the internal consistency of choice and matching tasks. That is, we test for preference reversals *within* methods, whereas the usual preference reversal is *between* methods. If preference reversals are also found within methods, this seems to be an even more elementary violation.

Our test relies on the observation that in tasks with two options, each consisting of two attributes, there are actually two ways to perform the matching task, and two ways to perform the choice task. These four different procedures are all *strategically equivalent*, i.e., they should generate the same preference orderings according to standard economic theory. Now, given the finding of systematically different results *between* choice and matching procedures,

---

frequency among pricing than among choices.

we can investigate the results of the two procedures *within* matching and the results of the two procedures *within* choice. This allows us to test whether choice or matching generates more inconsistencies (i.e., preference reversals) across different variations. If the choice task for example generates the same results for its two strategically equivalent variants, whereas the matching task generates systematic differences, this would suggest that the former has a higher internal consistency than the latter. It is known, for instance, that the two different matching procedures cause significantly different results (Delquié, 1993).

The main contribution of this study is that it provides a criterion (internal consistency) that can be used to assess the relative performance of different elicitation methods. Although this is of course merely one out of several possible evaluation criteria, it is a first step in developing these criteria, which may help to develop more guidance in the choice between alternative elicitation methods.

We introduce terminology and explain underlying theory in Section 2. Section 3 describes the details of our experiment. The results of the experiment are presented in Section 4. Finally, Section 5 contains a discussion of the results and concludes this paper.

2. Background

2.1. Terminology and Notation

We consider preferences over sets of outcomes, with the preference relation $\succcurlyeq$ assumed to be a weak order, i.e., transitive (if $x \succcurlyeq y$ and $y \succcurlyeq z$, then $x \succcurlyeq z$) and complete (either $x \succcurlyeq y$, or $y \succcurlyeq x$, or both). The relation $\succcurlyeq$ is the commonly adopted preference relation in decision theory, with $\succ$ denoting the asymmetric part of $\succcurlyeq$ and $\sim$ denoting indifference (i.e., the symmetric part). We consider outcomes X, each consisting of two attributes x_1 and x_2, that is,

$X=(x_1, x_2)$. x_1 denotes a particular health state and x_2 denotes a number of periods in that health state. An individual is assumed to prefer more life years to less in any health state, i.e., the relation $\succcurlyeq$ satisfies monotonicity in duration: $(x_1, x_2) \succ (x_1, x_2')$ for all chronic health states (x_1, x_2), (x_1, x_2') with $x_2 > x_2'$. Monotonicity cannot be defined for health status, because this is a qualitative variable. Instead, we assume health status is preferentially independent, i.e., for all life durations unequal to zero and for all health states x_1, x_1', $(x_1, x_2) \succcurlyeq (x_1', x_2) \Leftrightarrow (x_1, x_2') \succcurlyeq (x_1', x_2')$ (Bleichrodt and Pinto, 2005). Preference relations over the attributes are derived from $\succcurlyeq$. We only investigate two options, $Y=(y_1, y_2)$ and $Z=(z_1, z_2)$, at a time, with always either $y_1 \succcurlyeq z_1$ and $y_2 \preccurlyeq z_2$, or $y_1 \preccurlyeq z_1$ and $y_2 \succcurlyeq z_2$. That is, one option never dominates another and tradeoffs have to be made.

In a choice task, an individual has to choose one option of the set {Y, Z}, with the values of all attributes given. In matching, one attribute y_i or z_i in one option Y or Z is left blank and the subject is asked to give a value for this attribute that makes him indifferent between the options. A *preference reversal* is defined as the case where an individual in one procedure indicates that $Y \succcurlyeq Z$, but reveals the opposite preference, i.e., $Y \preccurlyeq Z$, in the other procedure, with at least one preference strict.

The profile $h=(h_1,\ldots,h_T)$ denotes a *health profile* where h_t is the health state in period $t=1,\ldots,T$, with T the decision maker's final period of life. A *constant health profile* $h=(h_1=\alpha,\ldots,h_T=\alpha)$ is indicated, using the above format, as the option $X=(x_1=\alpha, x_2=T)$. In other words, a constant health profile is described as an option with two attributes, the quality of the health state and the duration of that health state. Further, $v(h_t)$ is a value function that represents the individual's preferences over health quality and $\delta(t)$ denotes the corresponding weight given to the value in this period. It can then be shown that, under some reasonable axioms, $h \succcurlyeq h'$ if and only if $\sum_{t=1}^{T}\left(\delta(t)v(h_t)\right) \geq \sum_{t=1}^{T}\left(\delta(t)v(h_t')\right)$ (Bleichrodt and Gafni, 1996).

Similarly, the preference relation Y≽Z can be evaluated by $\sum_{t=1}^{y_2}\left(\delta(t)v(y_{1t})\right) \geq \sum_{t=1}^{z_2}\left(\delta(t)v(z_{1t})\right)$.

We assume that health profiles are evaluated by the function $U(t,h_t) = \sum_{t=j}^{T}\left(\delta(t)v(h_t)\right)$. We term the function $\sum_{t=j}^{T}\delta(t)$ the *utility of life duration* for the period between t=j and t=T. For convenience, we adopt the following notation throughout:

(1) $$W(T) = \sum_{t=1}^{T}\delta(t),$$

where W(T) is the total utility given to the period between t=1 and t=T.

The TTO method infers health state utilities by asking subjects to consider two constant health profiles: B=(β, n_β) and G=(γ, n_γ) with, in general, γ>β and n_β>n_γ. When an individual is indifferent between these two profiles ((β, n_β)~(γ, n_γ)), our model evaluates these preferences by the equation:

(2) $$W(n_\beta)v(\beta) = W(n_\gamma)v(\gamma).$$

Then, if we normalize v(γ) to 1, we get the following expression for v(β):

(3) $$v(\beta) = \frac{W(n_\gamma)}{W(n_\beta)}.$$

Eq. 3 makes clear that, in addition to the values of n_γ and n_β (of which one is a stimulus), one also needs to elicit the utility of life duration function W(.) in order to estimate

$v(\beta)$. We included a separate task for this (see Section 3), but the focus of our study is on the procedures to elicit n_γ and n_β.

As indicated above, the equivalence between two options can be obtained via a matching task, where the subject is asked to give a value for one blank attribute that makes him indifferent between the options. Then, one can fix n_β and directly ask for a value of n_γ that makes the subject indifferent between the health profiles, or do the opposite, i.e., fixing n_γ and asking for n_β.[2]

Second, in a choice task, an individual has to choose one of the two options. However, one choice only reveals that $B \succcurlyeq G$ or $B \preccurlyeq G$, and more choices are needed to yield an indifference relation. Then, depending on this choice, the value of either n_β (if B [G] was chosen, n_β is decreased [increased]) or n_γ (if B [G] was chosen, n_γ is increased [decreased]) is varied.[3] Repeating this procedure allows the elicitation of the indifference point. We use the bisection method for this, an algorithm which repeatedly bisects an interval and then selects a subinterval in which an indifference point must lie for further processing. Subsequently, we determine the indifference point as the midpoint of this interval.

We therefore separate four procedures to elicit the indifference of interest. These procedures are strategically equivalent, so that in principle, they should return the same preference orderings. Suppose, for example, that we have two health profiles, described by the options $B=(\beta, n_\beta)$ and $G=(\gamma, n_\gamma)$, where β is back pain, n_β the number of years of life with back pain (after which death follows), γ is full health and n_γ the number of years of life in full health (after which again death follows). If we employ a matching procedure, we can now fix either n_β or n_γ, and ask for the value of the remaining attribute, such that this value renders the subject indifferent between these options. Suppose that we fix the value of n_β at 10 years and the subject states that indifference occurs at $n_\gamma=6$ in a matching task. This gives the

[2] Because the health states are qualitative, these are both fixed in all procedures.

indifference relation (β, 10)~(γ, 6). Procedural invariance entails that this indifference should hold no matter what procedure is used. If we, for example, fix n_γ=6 and ask for n_β for the same health states, we should obtain n_β=10 for this subject. Similarly, using a bisection choice procedure where n_β is held fixed should also lead to that indifference, just as setting n_γ=6 should result in n_β=10 after a sufficient number of choices.

In the above example, where the subject had indicated to be indifferent between (β, 10) and (γ, 6) in the matching task, and under the usual assumption that this subject prefers more life years to less and better quality of life to worse, she should choose B over G when n_γ<6 and G over B when n_γ>6 in a choice task. If this is not the case, this means that the subject may, in fact, not be indifferent at the stated value. If she for example chooses the option G=(γ, 5) over B=(β, 10), she thus reveals G≽B even though G has become less attractive compared to the situation at which she was indifferent between B and G in the matching task, while B did not change. In other words, there are values for this subject where B≻G according to the matching question, but at the same time G≽B according to the direct choice for G. Thus, this response pattern corresponds to a preference reversal.

For convenience, we suppress notation in what follows and simply denote $v(\beta)$ by v. We use the following terms to distinguish the estimates of v obtained by the four different procedures:

-*Fixed-n_β choice procedure*: v_β^c indicates the estimates obtained by the choice task while fixing the duration in health state β (n_β).

-*Fixed-n_γ choice procedure*: v_γ^c indicates the estimates obtained by the choice task while fixing the duration in health state γ (n_γ).

[3] Obviously, one can also change both n_β and n_γ simultaneously (Delquié, 1997), but we do not pursue this possibility here.

-*Fixed-n_β matching procedure*: v_β^m indicates the estimates obtained by the matching task while fixing the duration in health state β (n_β).

-*Fixed-n_γ matching procedure*: v_γ^m indicates the estimates obtained by the matching task while fixing the duration in health state γ (n_γ).

2.2. Method

<TABLE 1 HERE>

Table 1 represents our within-subject, 2-by-2, design. It makes clear that two factors are relevant. One factor is the elicitation method, choice or matching (the horizontal comparison). The other factor is the direction of the trade-off: giving up or gaining years of life (the vertical comparison). The horizontal comparison therefore tests for preference reversals between choice and matching. The vertical comparison, on the other hand, tests for the reversibility of the indifference curves obtained by the two procedures. This can be clarified by Figure 1. Suppose we employ the fixed-n_β procedure, fixing (β, n_β) and infer the amount of n_γ such that (β, n_β) ~ (FH, n_γ). This indifference could be represented by an indifference curve comparable to the solid line in Figure 1. (Of course, in reality we need more indifference values to know the precise shape of the curve.) Procedural invariance then requires that, if we apply the fixed-n_γ procedure, we can use the same indifference curve, i.e., starting from (FH, n_γ) and moving towards (β, n_β). Hence, the indifference curve would be reversible in that case. If, however, the respondent instead elicits the indifference relation (FH, n_γ) ~ (β, n_β'), with $n_\beta' > n_\beta$, this would imply another indifference curve, as shown by the dashed line, indicating irreversibility of the indifference curve (Bleichrodt et al., 2003).

<FIGURE 1 HERE>

Scale compatibility is relevant in the horizontal comparison of Table 1, because it is liable to affect matching but not choice. However, scale compatibility is not relevant for the vertical comparison within matching because there is no change of response scale between these two matching tasks, which is the number of life years in both. We therefore attribute any difference between v_{β}^{m} and v_{γ}^{m} to reference effects (e.g., loss aversion). Scale compatibility implies that life years get more weight than health status in the decision of the subject (Bleichrodt and Pinto, 2002). As a result, they are more resistant to give up life years in the fixed-n_{β} procedure, and demand fewer additional life years in the fixed-n_{γ} procedure to compensate for the deterioration in quality of life. Scale compatibility is therefore predicted to increase TTO scores in both procedures of the matching task (Bleichrodt, 2002).

Loss aversion reinforces the effect of scale compatibility in the fixed-n_{β} procedure, because the life years given up are considered a loss and get more weight, which results in even higher TTO scores. On the other hand, loss aversion reduces this upward tendency in the fixed-n_{γ} procedure, since the loss of health gets a penalty and the individuals demand extra life years in return. Because both loss aversion and scale compatibility are empirically well-established, it is therefore not clear, a priori, whether there is an upward or downward tendency on TTO scores in the fixed-n_{γ} procedure.

Loss aversion is most often modeled by taking a fixed reference point and multiplying the (utility of the) loss by some parameter $\lambda>1$. In terms of the two option-two attribute example, one option has a gain in one attribute relative to the other option, and a loss in the other attribute. The main issue then is what is taken as the reference point. It seems most natural that, in case of a matching task, the option where both attributes are given is taken as the reference point. Subsequently, the subject will compare the value of the known attribute of

the other option to its value in the reference scenario. If the former is lower than the latter, this is considered a loss and given more weight. The subject will consequently demand a higher gain in the other attribute in return. For instance, consider the fixed-n_β matching procedure. The values of both attributes of option B are given and it is likely that B is taken as the reference point. This is compared to G where the first attribute involves a gain ($\gamma > \beta$), which is traded off against the loss in the second attribute ($n_\gamma < n_\beta$). Because this loss gets more weight under loss aversion, the subject is more inclined to choose B and thus to demand extra n_γ to make G as attractive as B. The empirical findings of loss aversion (e.g., Kahneman et al., 1990; Knetsch, 1989; Tversky and Kahneman, 1991) provide highly relevant evidence for the latter.

2.3. Tests

Given the presented framework, our experimental design allows for a number of tests. First, we can replicate earlier tests for differences between matching and choice tasks, and for the reversibility of the indifference curves. Second, a novelty of our study is that we perform these tests simultaneously and, hence, are able to test differences between matching and choice tasks *for two different procedures*, and to test reversibility of the indifference curves *for both choice and matching*. This allows us to test whether, for example, choice tasks are more or less susceptible to irreversibility of the indifference curve than matching tasks. In particular, we test three hypotheses, which are presented in turn below.

2.3.1. Scale compatibility

The first hypothesis comprises the horizontal comparison of Table 1, i.e., whether choice and matching generate the same values, both for the fixed-n_β and the fixed-n_γ procedure (H_0: $v_\beta^c = v_\beta^m$ and $v_\gamma^c = v_\gamma^m$). Our framework predicts higher values for matching than

for choice due to scale compatibility, whereas loss aversion is expected to have the same influence for choice and matching. Therefore, we expect the choice task to generate lower TTO scores than the matching task, for both procedures (H_A: $v_\beta^c < v_\beta^m$ and $v_\gamma^c < v_\gamma^m$).

2.3.2. Loss aversion

The second hypothesis similarly involves the vertical comparison of Table 1, i.e., whether the fixed-n_β and the fixed-n_γ procedures generate the same values, both for the choice task and the matching task (H_0: $v_\beta^c = v_\gamma^c$ and $v_\beta^m = v_\gamma^m$). In our framework, loss aversion has an upward influence on TTO scores in the fixed-n_β procedure, and a downward influence on TTO scores in the fixed-n_γ procedure, for both tasks (Bleichrodt, 2002); whereas, as explained above, the influence of scale compatibility does not depend on the procedure. Therefore, we expect the fixed-n_β procedure to generate higher TTO scores than the fixed-n_γ procedure, for both tasks (H_A: $v_\beta^c > v_\gamma^c$ and $v_\beta^m > v_\gamma^m$).

2.3.3. Preference reversals within tasks

Our framework does not predict any difference in irreversibility of indifference curves between the two tasks (i.e., choice or matching, second hypothesis). However, Bostic et al. (1990) found that when prices for gambles are elicited by means of a series of choices instead of directly stating a particular amount (i.e., matching), fewer direct choice-pricing preference reversals are generated. Moreover, Schmidt and Hey (2004) found evidence for more errors in pricing than in choices. Since pricing can be thought of as a kind of matching, this also suggests more consistency within choices. If this empirical evidence is transferable to preference reversals caused by different procedure as tested in our second hypothesis, we may therefore expect the matching task to be more likely to generate preference reversals than the choice task. In order to test this more formally, we tested a third hypothesis. This hypothesis

tested whether internal consistency within choice was similar to internal consistency within matching. We tested this hypothesis by first computing the absolute differences between the TTO scores for the two procedures, both for the choice task and for the matching task. Therefore, we tested whether $\left|v_{\beta}^{c} - v_{\gamma}^{c}\right| = \left|v_{\beta}^{m} - v_{\gamma}^{m}\right|$. An extension of the findings by Bostic et al. (1990) to our context would imply that choice tasks produce less divergence between the fixed-n_{β} procedure and the fixed-n_{γ} procedure than matching tasks (H_A: $\left|v_{\beta}^{c} - v_{\gamma}^{c}\right| < \left|v_{\beta}^{m} - v_{\gamma}^{m}\right|$).

3. Experiment

3.1. Subjects

The experiment was performed with 80 undergraduate Business Administration students from Erasmus University Rotterdam. They received course credits for their participation. Because we used health outcomes, it was not possible to use real incentives.

3.2. Procedure

The experimental sessions were run by one of the authors with four subjects at a time. The subjects were separated by partitions, in order to avoid discussion between them. The sessions lasted 30 minutes on average. The experiment was fully computerized and entailed the four different TTO tasks, as well as the utility of life duration elicitation task. In addition, some other tasks that were part of another study were included in-between these tasks. This was expected to reduce remembrance effects. The TTO tasks were ordered in four different ways. In particular, there were four computers, each having another ordering (see Table 2). Subjects were allocated to these computers randomly. Besides controlling for ordering effects, this design enabled a robustness (between-subjects) test of preference reversals. This may be

relevant if preference reversals are found between-subjects, but not within-subjects, because this would suggest that recall effects abolish within-subjects preference reversals. The utility of life duration task was performed at the end of the experiment in all four versions. Practice questions were included at the beginning of each task.

<TABLE 2 HERE>

We described the health state using the domains contained in the EuroQol 5D (EQ-5D) questionnaire. The EQ-5D is a popular questionnaire for eliciting health state utilities (Dolan, 1997). It describes health states in terms of five dimensions, each consisting of three levels, which indicates how the subject is functioning on these dimensions. The five dimensions are mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The descriptions were printed on cards and handed to the subjects (see Appendix). It was made clear to the subjects that health state γ meant they were able to function perfectly on all five EQ-5D dimensions, irrespective of their age. The same health states were used in the utility of life duration and TTO parts to avoid distorting influences of different behavior for different health states (e.g., if different utility functions existed for different health states).

The choice [matching] task of the fixed-n_β procedure fixed the number of years with health problems and varied [asked for] the number of years in full health until [for which] the subject was indifferent. Conversely, the choice [matching] task of the fixed-n_γ procedure fixed the number of years in full health and varied [asked for] the number of years with health problems until [for which] the subject was indifferent.

The answers of the subject in the fixed-n_β [fixed-n_γ] procedure were stored and used as the gauge duration for that subject in the subsequently included fixed-n_γ [fixed-n_β] procedure. The rationale for this procedure was to reduce distorting influences, such as an influence of

utility of life duration. If a subject was internally consistent, her answer in the last procedure then had to be equal to the initial gauge duration used in the first procedure (Attema and Brouwer, 2008; Attema and Brouwer, forthcoming; Bleichrodt et al., 2003).

In order to check general data quality, we included a reliability test in the choice task by repeating the second or third iteration for some randomly selected questions at the end of an iteration sequence.

3.2.1. Utility of life duration

Utility of life duration was elicited by the risk-free direct method (Attema et al., forthcoming), also using a bisection procedure. The subjects' task in this method is to compare two different health profiles, each consisting of two health states: β and γ, with γ>β. In the first health profile, $A=(h_1=\gamma,\ldots,h_m=\gamma,h_{m+1}=\beta,\ldots,h_T=\beta)$, the subject gets an immediate improvement in health from β to γ, which lasts until time point *m*, after which the subject returns to health state β until point *T*. In the second health profile, $B=(h_1=\beta,\ldots,h_m=\beta,h_{m+1}=\gamma,\ldots,h_T=\gamma)$, he starts in health state β and will be in that health state until time point *m*. Then he will get the health improvement towards health state γ, and he will remain in γ until time point *T*. After *T*, the two options yield the same health state. Suppose that γ and β represent full health (i.e., EQ-5D state 11111) and EQ-5D state 21111 (i.e., a state with no problems except for some problems in walking about), respectively. Profile *A* means, in this case, that the subject will be in full health (and, hence, will have relief from problems in walking about) from now on during the following *m* years. After *m* years, the relief ends and he will have some problems in walking about from point *m* until point *T*. Making use of choices, we can find an indifference value for *m*, which means that the subject is indifferent between *A* and *B*. Considering the above framework, this indifference gives the following equality:

(4) $$W(m)v(\gamma)+[W(T)-W(m)]v(\beta)=W(m)v(\beta)+[W(T)-W(m)]v(\gamma).$$

Attema et al. (forthcoming) show that estimates of $W(m)=\frac{1}{2}W(T)$ can be obtained in this way. $W(T)$ can be normalized to 1, so that we in fact elicit a midpoint, which splits the total period into two parts (one before and one after) that both give the same utility of life duration. For example, if m=20 and T=50, this implies that the first 20 years are given the same value as the 30 years thereafter. More details about the implementation of this method can be found in Attema and Brouwer (2009).

The utility of life duration task also included a reliability test, by repeating the first question of the first iteration at the end of that iteration. The answer given in the initial question was used in the analysis, so the answer to the repeat question was solely used for the reliability test.

3.3. Stimuli

We considered two constant health profiles throughout the experiment, denoted by the two-attribute options B=(β, n_β) and G=(γ, n_γ), with β representing 'EQ-5D state 21111' and γ representing 'full health'. The health state '21111' is a mild health state and subjects were likely to know people suffering from it.

The values of the gauge durations depended on the version (Table 2). For example, in version 1, we could set the gauge durations of the fixed-n_β choice procedure and the fixed-n_γ matching procedure. However, the gauge durations for the fixed-n_γ choice procedure and the fixed-n_β matching procedure were the answers given by the subjects in the former two procedures, and, hence, varied between the subjects. We used the following gauge durations in the fixed-n_β procedure (both for the choice and the matching task) for the subjects who

received the fixed-n_β procedure first (versions 1 and 4 for choice, and versions 2 and 3 for matching): n_β=3, 10, and 31 years. For example, in the fixed-n_β procedure with gauge duration n_β=3, the subject was asked to imagine living 3 more years in state 21111 and then die. The alternative was to regain full health, at the expense of lifetime. The subject was then asked to state the number of years in full health such that the two options were indifferent (matching), or to choose the baseline scenario (21111, 3 years) or its alternative (11111, x years) (choice).

We also included three fixed-n_γ questions both for choice and for matching. The subjects answering the fixed-n_γ procedure first (versions 2 and 3 for choice, and 1 and 4 for matching) got the following gauge duration: n_γ^m=2, 7, and 26 years. Now, the remainder of the sample got a gauge duration which was the answer they gave in the fixed-n_β procedure. For ease of comparison, we will only refer to n=3, 10, and 31 years in the remainder of this paper. Hence, when writing n=31, we can also refer to n_γ^m = 26, or the answer given to n_γ^m = 26. This will be clear from the context.

The first question in both procedures of the choice task was intended to test whether subjects had understood the task: it used the same duration for both health states and, hence, the option with full health was dominant. If a subject chose for the option with inferior health, this implied a violation of monotonicity because a life of given duration in inferior health was preferred to a life of the same length in full health. The subject was given an explanation for why this choice may not be very reasonable and given the opportunity to reconsider her choice. If the subject still held on to this choice, she was redirected toward the next question. These subjects were not included in the analysis. A similar procedure was used for the matching task. There, a subject was given an explanation when her answer implied one of the

options was strictly better in both attributes, so that this could not represent a reasonable indifference.

In the matching task, the subject could give his answers up to two decimals. The minimum in the fixed-n_β procedure was 0, while there was no maximum in the fixed-n_γ procedure. For choices, the minimum in the fixed-n_β procedure could not reach exactly 0. It was 0.1 for n=3, 0.25 for n=10, and 1 for n=31. The maximum for the fixed-n_γ procedure was 64.5 for n=2, 68.5 for n=7, and 77 for n=26.[4] The constraints for the procedure that the subject received at second depended on the answers the subject gave in the procedure received at first and, hence, differed between subjects. Because we used a very mild health state, we did not allow subjects to value a health state negatively (i.e., as worse than dead), which would require a separate procedure (Torrance, 1986). However, this seems not to have biased our results, since none of the subjects attached the lowest possible value to this health state in the fixed-n_β procedure.

3.4. Analyses

The hypotheses were tested by comparing the TTO scores (v), since different TTO scores captured the response patterns that caused preference reversals. In particular, we compared the estimates of v for the four different procedures. Only nonparametric Wilcoxon signed ranks tests are reported for the within-subjects analyses, since the data were nonnormally distributed for 5 of the 12 variables (Kolmogorov-Smirnov test, $p<0.05$). In case the conclusion implied by the paired t-test differed from that implied by the Wilcoxon test, we also report the results of the former.

The between-subjects analysis only included the answers given to the *first two procedures* the subject received. Consequently, this only involved questions with exogenous

gauge durations, eliminating heterogeneity in the stimuli. Moreover, there is no possible distortion caused by recall in this analysis. The experiment was designed in such a way that all hypotheses could be tested with similar power. Hypothesis 1 was tested by comparing v_β^c for subjects in versions 1 and 4 to v_β^m for subjects in versions 2 and 3, and v_γ^c for versions 2 and 3 was compared to v_γ^m for versions 1 and 4. Hypothesis 2 was tested by comparing v_β^c for versions 1 and 4 to v_γ^c for versions 2 and 3, and v_β^m for versions 2 and 3 to v_γ^m for versions 1 and 4. These analyses were performed by means of nonparametric Mann-Whitney tests. In case the conclusion implied by the independent samples t-test differed from that implied by the Mann-Whitney test, we also report the results of the former. All reported p-values are two-sided.

4. Results

We removed the data of four subjects from the analysis, because their answers violated dominance. Therefore, 76 subjects (mean age 19.9, s.d. 1.8, 48 men) were included in the analysis. Table 3 shows the distribution of the subjects across the four versions.

<TABLE 3 HERE>

The replication of the second iteration led to the same choice as the initial question in 98% of the cases, indicating a good reliability. The replication of the third iteration led to the same choice in 77.6% of the cases. This lower reliability was to be expected, since the

[4] The minimum answer given in the fixed-n_β procedure of the matching task was 1, and the maximum answer given in the fixed-n_γ procedure was 80. Therefore, the lack of a constraint did not distort the comparison with the

stimulus value was likely to be closer to the indifference value in the third iteration. Reliability in the discounting task was also good, with 90.8% choosing the same option in the replicated question.

4.1. Choice vs. Matching

Figures 2 compares the medians of the TTO scores obtained by the fixed-n_β choice and the fixed-n_β matching procedures, plotted against the gauge durations. Figure 3 does the same for the fixed-n_γ procedures. The figures indicate a violation of procedural invariance. In particular, the results make clear that the matching task indeed generated higher TTO scores over the entire duration range for the fixed-n_β procedure, although the difference was not very large and declined with duration. For the fixed-n_γ procedure, however, the TTO scores were significantly *higher* for choice than for matching, with the difference between tasks substantially greater than for the fixed-n_β procedure. The only resemblance to the fixed-n_β procedure was the declining difference for longer durations.

<FIGURE 2 AND 3 HERE>

Tables 4 and 5 give more detailed statistics. Concerning the fixed-n_β procedure, the TTO scores differed significantly between choice and matching for two out of the three gauge durations ($p<0.01$ for n=3 and 10; $p=0.08$ for n=31).[5] Furthermore, the TTO scores were significantly higher for choice than for matching in the fixed-n_γ procedure, for all three gauge durations ($p<0.01$), implying a rejection of Hypothesis 1 for this procedure as well. However, this rejection was not in favor of the alternative hypothesis, which predicted higher TTO

choice task.

[5] The p-values for the paired t-test were $p<0.01$, $p=0.10$, and $p=0.13$ for the n=3, 10, and 31, respectively.

scores for matching than for choice due to scale compatibility being expected to exert an upward influence in matching but not in choice.

<TABLES 4 AND 5 HERE>

Our results were confirmed in the between-subjects test. The Mann-Whitney test indicated higher TTO scores for matching than for choices in the fixed-n_β procedure, but the difference was significant at the 5% level only for n=3 (p=0.02)[6]. The opposite pattern occurred for the fixed-n_γ procedure, with TTO scores for matching being significantly lower than TTO scores for choices ($p<0.05$ for all durations)[7].

4.2. Procedural invariance

Figures 4 and 5 compare the two variants within the matching task and within the choice task, respectively, again plotted against the gauge durations. These are the same data points as those of Figures 2 and 3, but combined in a different way. Interestingly, the matching task generated a substantial difference between the fixed-n_β and the fixed-n_γ procedures, especially for short durations. The choice task, by contrast, did not generate these differences.

<FIGURES 4 AND 5 HERE>

[6] According to the t-test, this difference was significant at the 6% level (p=0.06 for n=3).
[7] The difference was not significant for n=31 according to the t-test, however (p=0.17).

Table 6 shows the relevant differences for Hypothesis 2, as computed from the data given in Tables 4 and 5. The null of Hypothesis 2 could be rejected, in favor of the alternative, in the matching task ($p<0.01$ for all three gauge durations). That is, the fixed-n_β procedure generated higher TTO scores than the fixed-n_γ procedure. However, the evidence was not so clear for the choice task. There we found significantly higher TTO scores under the fixed-n_β procedure for the 3-year gauge duration ($p<0.01$), no significant difference for the 10-year gauge duration ($p=0.36$), and significantly lower TTO scores under the fixed-n_γ for the 31-year gauge duration ($p<0.05$)[8].

<TABLE 6 HERE>

Between subjects, TTO scores were significantly higher under the fixed-n_β matching procedure than under the fixed-n_γ matching procedure for n=3 and n=10 ($p<0.01$), and marginally significantly higher for n=31 ($p=0.06$)[9]. Within the choice task, no significant differences were found between the two procedures in the between-subjects test ($p>0.08$). Hence, the lack of significant differences within the fixed-n_β procedure and within choices in the within-subjects analysis did not seem to be due to remembrance effects, since the between-subjects analysis gave the same conclusions, and even showed a tendency towards more equality within the choice task.

4.3. Preference reversals within choice and within matching

The results of Section 4.2 suggest a higher rate of preference reversal within matching tasks than within choice tasks. The third hypothesis could test this suspicion more formally. This required the computation of the absolute values of the within-task differences between

[8] However, this difference is only significant at the 7% level in the paired t-test.
[9] $p=0.44$ according to the t-test.

TTO scores generated by the fixed-n_β and fixed-n_γ procedures. Table 7 presents some resulting summary statistics.

<TABLE 7 HERE>

The statistical tests confirmed our presumption. The absolute differences were higher within matching than within choice for all three gauge durations, and significantly so for two of them ($p<0.01$ for $n_\beta = 3(A2) / n_\gamma = 2(A3)$, and $n_\beta = 10(A7) / n_\gamma = 7(A10)$; $p=0.17$ for $n_\beta = 31(A26) / n_\gamma = 26(A31)$).

5. Discussion

The preference reversal phenomenon is a serious problem in decision theory. When strategically equivalent elicitation procedures produce different preference orderings, it becomes difficult to set priorities in budget allocations. The important question then arises which procedure should be used to generate the values to be used in these allocations. Previous studies on preference reversals mainly highlighted that different elicitation procedures generated systematically different preferences, but could not give directions as to whether one procedure was more valid than another. This paper has made an attempt to provide such directions, by introducing a new test for the internal consistency of different preference elicitation methods. In particular, we investigated whether differences exist in the degree to which different procedures capture people's preferences consistently. The main advantage of our test is that it provides a benchmark against which to evaluate the validity of different estimates, obtained from different elicitation procedures; a lower degree of

preference reversals for different procedures within one task than within another task would seem a logical reason to attach more validity to the former, assuming no other information is available.

Our study confirmed the pattern found in other studies on preference reversals. In particular, choice and matching elicitation procedures caused significant differences in valuations. In addition, we have confirmed the findings of Bostic et al. (1990) that choice tasks generate fewer inconsistencies than matching tasks. Bostic et al. (1990) used a series of choices to determine a cash amount that was indifferent to a given bet. They found that this led to fewer 'traditional' preference reversals than a matching task where the subjects simply had to state a cash amount that caused indifference. We found that eliciting indifference through choice series also reduces another kind of preference reversal, i.e., one that is caused by using different response modes within a task. Moreover, we used other outcomes (health instead of money) than Bostic et al. (1990), indicating that this pattern holds in other domains as well, and may therefore be a more universal phenomenon.

We also obtained some more surprising results. First, the finding of lower TTO scores for matching than for choice in the fixed-n_γ procedure is not in agreement with the result predicted by scale compatibility (Bleichrodt, 2002). Our framework offers two possible explanations for this finding. One possibility is that scale compatibility is not only present in the matching task, but also, and even to a higher extent, in the choice task. This explanation does not seem intuitively appealing, however, because the response scale is not likely to be more prominent in the choice task than in the matching task. A second explanation is that the deflating effect of loss aversion is stronger in the matching task than in the choice task (Attema and Brouwer, forthcoming; Bleichrodt et al., 2003). We believe that the latter explanation is more realistic, as the matching task puts more emphasis on the fact that something has to be *given up* in order to improve health status.

Another remarkable result is the finding of lower TTO scores for the fixed-n_β procedure than for the fixed-n_γ procedure for the longest gauge duration of the choice task. However, this difference was only significant at the 5%-level according to the Wilcoxon signed ranks test, and it was not significantly different according to the paired t-test. Hence, the difference seems not robust enough to attach any firm conclusion to this result.

Although not the main objective of this study, an interesting finding was the declining difference between procedures, as well as between tasks, for longer gauge durations. This suggests that the influence of loss aversion wears off with duration, which was also found in several other studies (Attema and Brouwer, 2008; Attema and Brouwer, forthcoming; Bleichrodt and Pinto, 2002; Bleichrodt et al., 2003). Hence, the use of longer gauge durations may be useful to reduce any bias due to loss aversion.

Some limitations of the present study are worth mentioning. First, we used a student population, which may hamper the generalization of the results. However, we have no reason to suspect that preference reversals occur less often among the general population. University students are likely to have higher cognitive skills than the average of the general population, and, therefore, could be expected to be less deviant from rationality.

Second, subjects may have recalled their answers in the choice [matching] procedure and put in those values in the matching [choice] procedure because they for example thought the experimenter wanted this. This reasoning is unlikely to have caused problems, however, since the matching and choice tasks were interspersed by filler tasks. Moreover, our findings of substantial differences between the procedures indicate that recall did not play a significant role. If it did, it would make the problem of preference reversals even more severe, because it would have decreased the frequency of preference reversals and, hence, this frequency would be even higher if there were no recall. The likely absence of recall effects is emphasized by the similar number of preference reversals in the between-subjects analysis.

Related to this second limitation, notice that subjects did not see the exact value of the point of indifference derived under the choice task (as this was inferred from their choices), but did for matching. Therefore, recall seems more likely in matching. The fact that we, nonetheless, observed larger differences for matching, further disqualifies matching.

A third limitation is that the implementation of real incentives was hindered by our use of health instead of monetary outcomes. However, there is no clear-cut evidence that real incentives generate different results than hypothetical incentives (Abdellaoui et al., 2011; Anderson and Mellor, 2009; Beattie and Loomes, 1997), only that they may reduce noise (Camerer and Hogarth, 1999).

The results we reported in this paper are of relevance not only for (experimental) economists, but for policy makers as well. For instance, budget allocation decisions in health care are often influenced by the outcomes of economic evaluations, which in turn rely on stated preference methods, such as the ones investigated in this paper, to estimate utilities resulting from different health interventions. Our results suggest these estimates are systematically and substantially different for alternative, strategically equivalent, procedures. Therefore, the choice of the elicitation procedure in an economic evaluation can have profound implications for priority setting in health care and the allocation of other public resources.

To summarize, we have shown that *within* choices there is more consistency between different procedures than *within* matching. Therefore, the neutral frame and closed-ended format of choice tasks seem to be less susceptible to biases than matching tasks. In that sense, the increase in questions necessary to elicit values in choice elicitation tasks may be considered worthwhile given its higher internal consistency. Our results give tentative support for a more widespread use of choice-based elicitation tasks in utility assessment. We emphasize, however, that a higher internal consistency is not synonymous to a better ability to

capture true preferences. Internal consistency is just a necessary, but not sufficient, characteristic of preference elicitation methods. More studies in this important field would be welcomed to confirm these findings and extend them to other characteristics – or to falsify them and, hence, reverse our preference for choice-based tasks.

Appendix: Health state descriptions

Card 1 – EQ-5D state 21111

You have regular back pain. This has the following consequences for your functioning in daily life:

- You have *some* problems in walking about.
- You have no problems to wash or dress yourself.
- You have no problems with your usual activities.
- You have no pain or other discomfort.
- You are not anxious or depressed.

Card 2 – EQ-5D state 11111 (Full Health)

You have no complaints and are in full health. This has the following consequences for your functioning in daily life:

- You have no problems in walking about.
- You have no problems to wash or dress yourself.
- You have no problems with your usual activities.
- You have no pain or other discomfort.
- You are not anxious or depressed.

Acknowledgments

Han Bleichrodt and Peter P. Wakker gave many helpful comments on a previous version of this paper. Participants of the Behavioral Economics Conference at the Erasmus School of Economics and participants of the Decision & Uncertainty Workshop at HEC are acknowledged for useful suggestions. The usual disclaimer applies.

References

-Abdellaoui M, L'Haridon O, Paraschiv C. Experienced vs. Described Uncertainty: Do We Need Two Prospect Theory Specifications? Management Science 2011;57; 1879-1895.

-Anderson L, Mellor J. Are risk preferences stable? Comparing an experimental measure with a validated survey-based measure. Journal of Risk and Uncertainty 2009;39; 137-160.

-Arrow KJ, Solow RM, Portney PR, Leamer EE, Radner R, Schuman H. Report of the NOAA Panel on contingent valuation. Federal Register 1993;58; 4601-4614.

-Attema AE, Bleichrodt H, Wakker PP. A direct method for measuring discounting and QALYs more easily and reliably. Medical Decision Making forthcoming;.

-Attema AE, Brouwer WBF. Can we fix it? Yes we can! But what? A new test of procedural invariance in TTO-measurement. Health Economics 2008;17; 877-885.

-Attema AE, Brouwer WBF. The correction of TTO-scores for utility curvature using a risk-free utility elicitation method. Journal of Health Economics 2009;28; 234-243.

-Attema AE, Brouwer WBF. The way that you do it? An elaborate test of procedural invariance of TTO, using a choice-based design. The European Journal of Health Economics forthcoming;.

-Beattie J, Loomes GC. The Impact of Incentives Upon Risky Choice Experiments. Journal of Risk and Uncertainty 1997;14; 155-168.

-Bleichrodt H, Pinto JL. The validity of QALYs under non-expected utility. The Economic Journal 2005;115; 533-550.

-Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Economics 2002;11; 447-456.

-Bleichrodt H, Gafni A. Time preference, the discounted utility model and health. Journal of Health Economics 1996;15; 49-66.

-Bleichrodt H, Pinto JL. Loss aversion and scale compatibility in two-attribute trade-offs. Journal of Mathematical Psychology 2002;46; 315-337.

-Bleichrodt H, Pinto JL, Abellán-Perpinán JM. A consistency test of the time trade-off. Journal of Health Economics 2003;22; 1037-1052.

-Borcherding K, Eppel T, von Winterfeldt D. Comparison of Weighting Judgments in Multiattribute Utility Measurement. Management Science 1991;37; 1603-1619.

-Bostic R, Herrnstein RJ, Luce RD. The effect on the preference-reversal phenomenon of using choice indifferences. Journal of Economic Behavior & Organization 1990;13; 193-212.

-Butler DJ, Loomes GC. Imprecision as an account of the preference reversal phenomenon. American Economic Review 2007;97; 277-297.

-Camerer CF, Hogarth RM. The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. Journal of Risk and Uncertainty 1999;19; 7-42.

-Cubitt RP, Munro A, Starmer C. Testing explanations of preference reversal. Economic Journal 2004;114; 709-726.

-Delquié P. Inconsistent trade-offs between attributes: new evidence in preference assessment biases. Management Science 1993;39; 1382-1395.

-Delquié P. "Bi-Matching": A new preference assessment method to reduce compatibility effects. Management Science 1997;43; 640-658.

-Dolan P. Modeling valuations for EuroQol health states. Medical Care 1997;35; 1095-1108.

-Dolan P, Gudex C, Kind P, Williams A. The time trade-off method: Results from a general population study. Health Economics 1996a;5; 141-154.

-Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. Journal of Health Economics 1996b;15; 209-231.

-Fishburn PC. SSB utility theory and decision-making under uncertainty. Mathematical Social Sciences 1984;8; 253-285.

-Fishburn PC. Nontransitive preference theory and the preference reversal phenomenon. Rivista Internazionale Di Scienze Economiche e Commerciali 1985;32; 39-50.

-Holt CA. Preference Reversals and the Independence Axiom. American Economic Review 1986;76; 508-515.

-Huber J, Ariely D, Fischer G. Expressing Preferences in a Principal-Agent Task: A Comparison of Choice, Rating, and Matching. Organizational Behavior and Human Decision Processes 2002;87; 66-90(25).

-Kahneman D, Knetsch JL, Thaler RH. Experimental Tests of the Endowment Effect and the Coase Theorem. Journal of Political Economy 1990;98 6; 1325-1348.

-Karni E, Safra Z. "Preference Reversal" and the Observability of Preferences by Experimental Methods. Econometrica 1987;55; 675-685.

-Knetsch JL. The Endowment Effect and Evidence of Nonreversible Indifference Curves. American Economic Review 1989;79; 1277-1284.

-Lamers LM, McDonnell J, Stalmeier PFM, Krabbe PFM, Busschbach JJV. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Economics 2006;15; 1121-1132.

-Lichtenstein S, Slovic P. Reversals of preference between bids and choices in gambling decisions. Journal of Experimental Psychology 1971;89; 46-55.

-Lindman HR. Inconsistent Preferences Among Gambles. Journal of Experimental Psychology 1971;89; 390-397.

-Loomes GC. Different experimental procedures for obtaining valuations of risky actions: Implications for utility theory. Theory and Decision 1988;25; 1-23.

-Loomes G, Starmer C, Sugden R. Preference Reversal: Information-Processing Effect or Rational Non-transitive Choice? Economic Journal 1989;99; 140-151.

-Loomes G, Sugden R. Regret Theory: An Alternative Theory of Rational Choice under Uncertainty. Economic Journal 1982;92; 805-824.

-Loomes G, Sugden R. A Rationale for Preference Reversal. American Economic Review 1983;73; 428-432.

-Maafi H. Preference Reversals Under Ambiguity. Management Science 2011;57; 2054-2066.

-Schmidt U, Hey JD. Are Preference Reversals Errors? An Experimental Investigation. Journal of Risk and Uncertainty 2004;29; 207-218.

-Seidl C. Preference Reversal. Journal of Economic Surveys 2002;16; 621-655.

-Stalmeier PFM, Wakker PP, Bezembinder TGG. Preference reversals: Violations of unidimensional procedure invariance. Journal of Experimental Psychology: Human Perception and Performance 1997;23; 1196-1205.

-Torrance GW. Measurement of health state utilities for economic appraisal. Journal of Health Economics 1986;5; 1-30.

-Tversky A, Thaler RH. Anomalies: preference reversals. Journal of Economic Perspectives 1990;4; 201-211.

-Tversky A, Kahneman D. Loss aversion in riskless choice: A reference-dependent model. Quarterly Journal of Economics 1991;106; 1039-1061.

-Tversky A, Kahneman D. Advances in Prospect Theory: Cumulative Representation of Uncertainty. Journal of Risk and Uncertainty 1992;5; 297-323.

-Tversky A, Sattath S, Slovic P. Contingent weighting in judgment and choice. Psychological Review 1988;95; 371-384.

-Tversky A, Slovic P, Kahneman D. The Causes of Preference Reversal. American Economic Review 1990;80; 204-217.

TABLES

TABLE 1. EXPERIMENTAL DESIGN

		Elicitation task: Choice	Elicitation task: Matching
Direction of trade-off	Fixed n_β (give up n)	v_β^c	v_β^m
	Fixed n_γ (gain n)	v_γ^c	v_γ^m

TABLE 2. ORDER OF THE TASKS

TASK NUMBER	Version 1	Version 2	Version 3	Version 4
1.	v_β^c	v_β^m	v_γ^c	v_γ^m
2.	Filler task	Filler task	Filler task	Filler task
3.	v_γ^c	v_γ^m	v_β^m	v_β^c
4.	Filler task	Filler task	Filler task	Filler task
5.	v_γ^m	v_γ^c	v_β^c	v_β^m
6.	Filler task	Filler task	Filler task	Filler task
7.	v_β^m	v_β^c	v_γ^m	v_γ^c
8.	Discounting	Discounting	Discounting	Discounting

TABLE 3. DISTRIBUTION OF SUBJECTS ACROSS VERSIONS [INCLUDED SUBJECTS WITHIN BRACKETS]

Version 1	Version 2	Version 3	Version 4
19 [18]	20	22 [20]	19 [18]

TABLE 4. FIXED-N_β PROCEDURE

		$n_\beta = 3/A(2)$	$n_\beta = 10/A(7)$	$n_\beta = 31/A(26)$
v_β^c	**Average**	0.74	0.76	0.82
	Std. deviation	0.21	0.18	0.17
	Median	0.78	0.80	0.87
	Interquartile range	0.61-0.91	0.65-0.90	0.78-0.93
v_β^m	**Average**	0.81	0.80	0.85
	Std. deviation	0.17	0.19	0.16
	Median	0.83	0.84	0.89
	Interquartile range	0.67-0.95	0.76-0.91	0.83-0.96
$v_\beta^m - v_\beta^c$	**Average**	0.07	0.04	0.03
	Median	0.05	0.04	0.02

TABLE 5. FIXED-N_γ PROCEDURE

		$n_\gamma = 2/A(3)$	$n_\gamma = 7/A(10)$	$n_\gamma = 26/A(31)$
v_γ^c	**Average**	0.66	0.75	0.85
	Std. deviation	0.22	0.18	0.13
	Median	0.69	0.80	0.90
	Interquartile range	0.53-0.83	0.66-0.87	0.80-0.94
v_γ^m	**Average**	0.49	0.62	0.79
	Std. deviation	0.25	0.20	0.16
	Median	0.45	0.62	0.85
	Interquartile range	0.29-0.67	0.50-0.78	0.72-0.90
$v_\gamma^m - v_\gamma^c$	**Average**	-0.17	-0.13	-0.06
	Median	-0.24	-0.18	-0.05

TABLE 6. COMPARISON OF PROCEDURES

		$n_\beta = 3 / A(2)$ $n_\gamma = 2 / A(3)$	$n_\beta = 10 / A(7)$ $n_\gamma = 7 / A(10)$	$n_\beta = 31 / A(26)$ $n_\gamma = 26 / A(31)$
v_β^m - v_γ^m	**Average**	0.32	0.18	0.06
	Median	0.38	0.22	0.04
v_β^c - v_γ^c	**Average**	0.08	0.01	-0.03
	Median	0.09	0	-0.03

TABLE 7. WITHIN-CHOICE AND WITHIN-MATCHING ABSOLUTE DIFFERENCES

		$n_\beta = 3 / A(2)$ $n_\gamma = 2 / A(3)$	$n_\beta = 10 / A(7)$ $n_\gamma = 7 / A(10)$	$n_\beta = 31 / A(26)$ $n_\gamma = 26 / A(31)$
$\left\|v_\beta^c - v_\gamma^c\right\|$	**Average**	0.15	0.10	0.08
	Std. deviation	0.15	0.13	0.11
	Median	0.11	0.05	0.04
	Interquartile range	0.03-0.25	0.03-0.12	0.02-0.10
$\left\|v_\beta^m - v_\gamma^m\right\|$	**Average**	0.33	0.20	0.10
	Std. deviation	0.24	0.18	0.11
	Median	0.31	0.17	0.06
	Interquartile range	0.14-0.49	0.08-0.27	0.02-0.12

FIGURES

FIGURE 1. REVERSIBILITY OF THE INDIFFERENCE CURVE

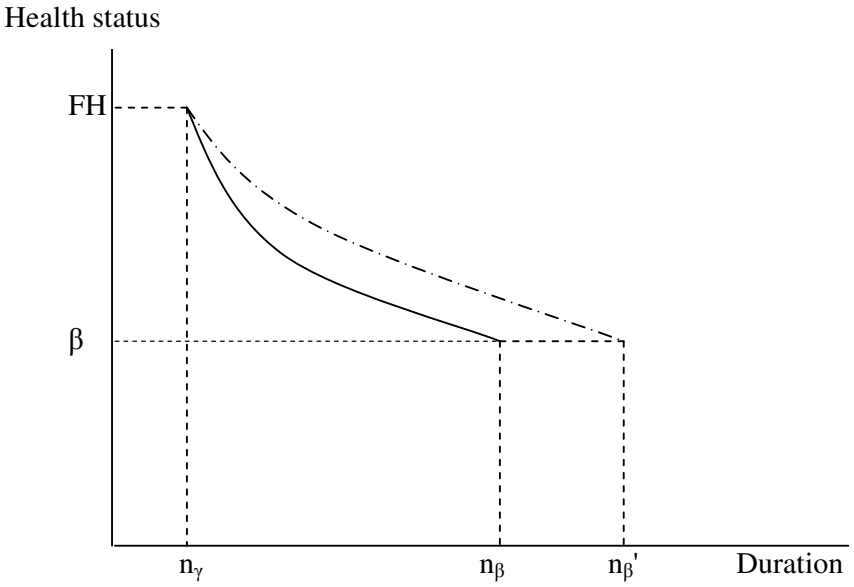

FIGURE 2. MEDIAN TTO SCORES FOR FIXED-N_β CHOICE AND MATCHING PROCEDURES, USING SEVERAL GAUGE DURATIONS

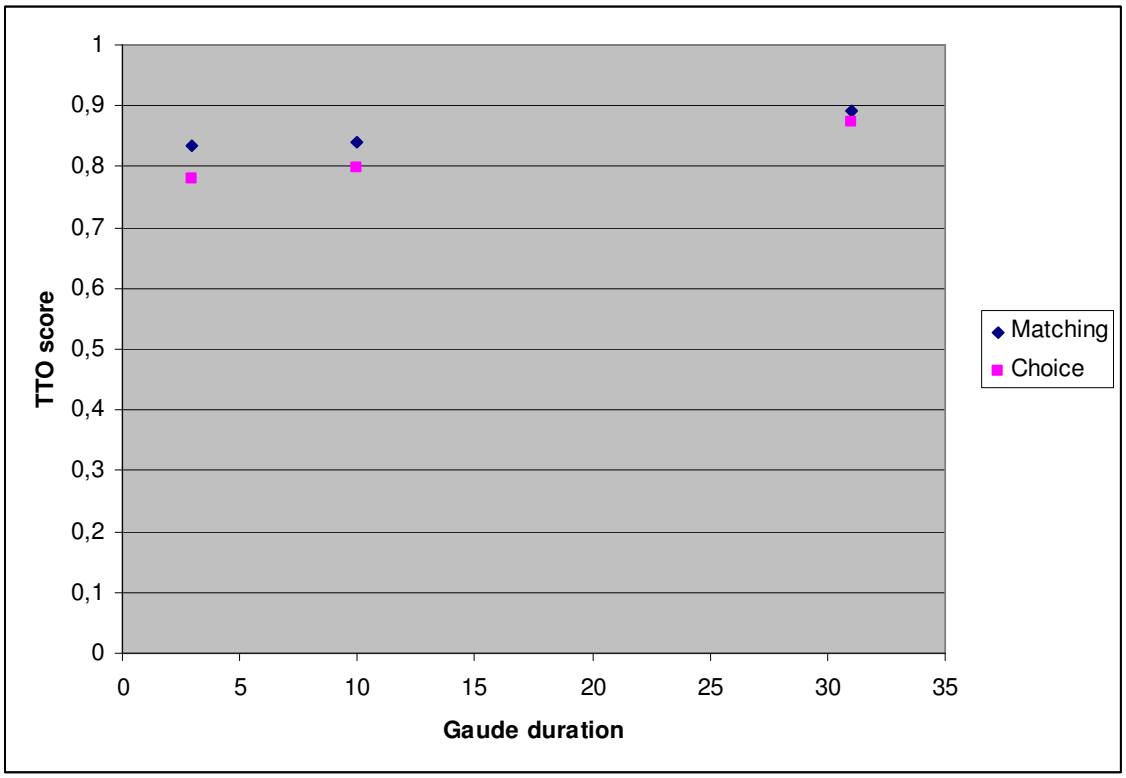

FIGURE 3. MEDIAN TTO SCORES FOR FIXED-N_γ CHOICE AND MATCHING PROCEDURES, USING SEVERAL GAUGE DURATIONS

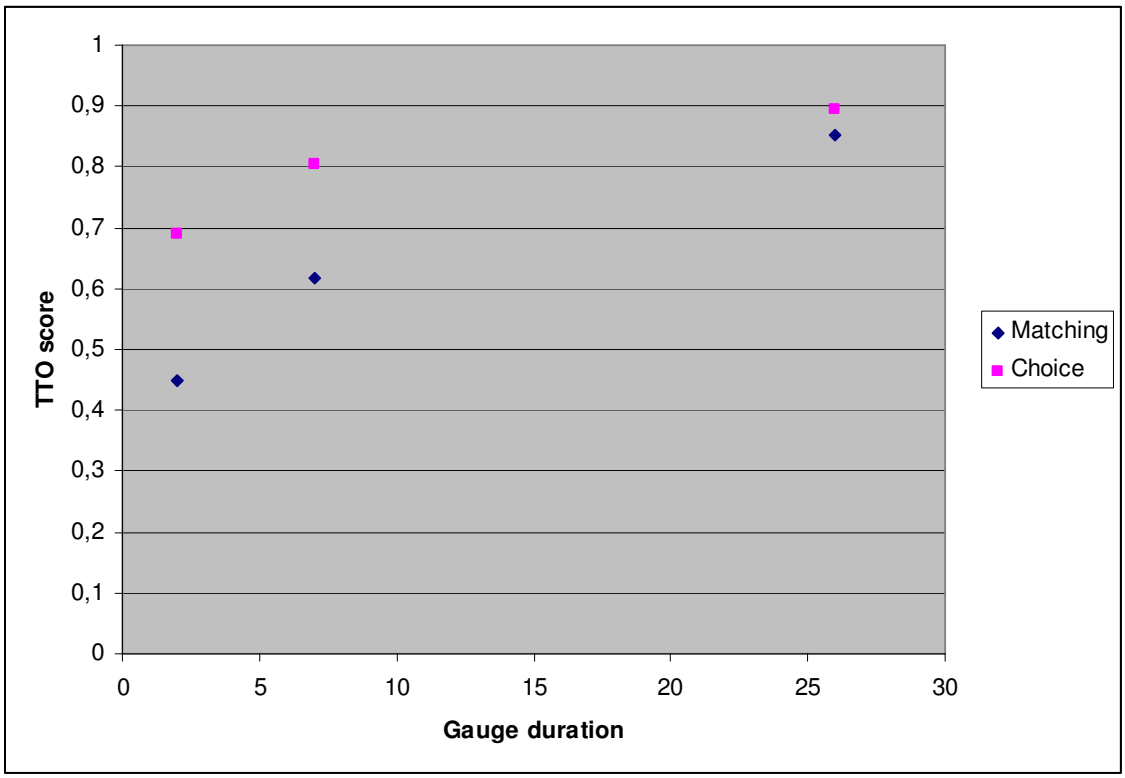

FIGURE 4. MEDIAN TTO SCORES FOR FIXED-N_β AND FIXED-N_γ MATCHING PROCEDURES, USING SEVERAL GAUGE DURATIONS

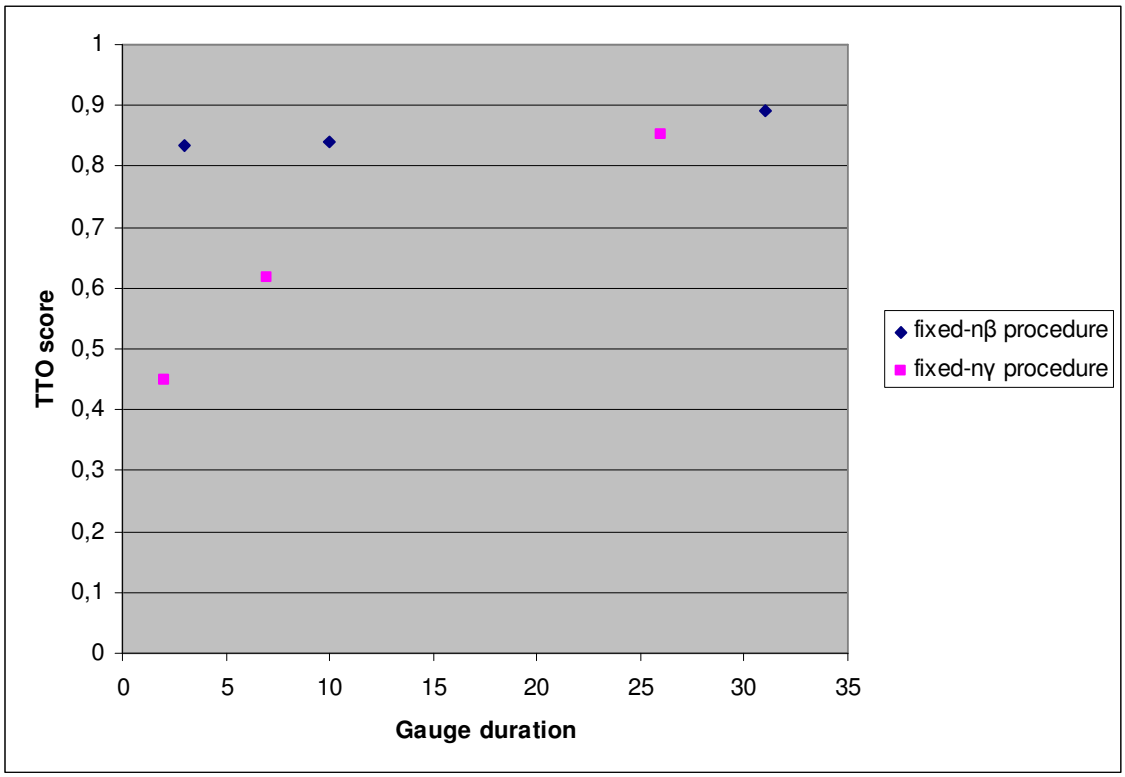

FIGURE 5. MEDIAN TTO SCORES FOR FIXED-N_β AND FIXED-N_γ CHOICE PROCEDURES, USING SEVERAL GAUGE DURATIONS