Journal of Experimental Psychology:
Human Perception and Performance
1997, Vol. 23, No. 4, 1196–1205

0096-1523/97/$3.00

Preference Reversals: Violations of Unidimensional Procedure Invariance

Peep F. M. Stalmeier
University of Nijmegen, St. Radboud Academic Hospital, and Amsterdam Academic Medical Centre

Peter P. Wakker
University of Leiden

Thom G. G. Bezembinder
University of Nijmegen

Preference reversals have usually been explained by weighted additive models, in which different tasks give rise to different importance weights for the stimulus attributes, resulting in contradictory trade-offs. This article presents a preference reversal of a more extreme nature. Let (10, 5 Migr) denote living 10 years with a migraine for 5 days per week. Many participants preferred (10, 5 Migr) to (20, 5 Migr). However, when asked to equate these two options with a shorter period of good health, they usually demanded more healthy life years for (20, 5 Migr) than for (10, 5 Migr). This preference reversal within a single dimension cannot be explained by different importance weights and suggests irrationalities at a more fundamental level. Most participants did not change their responses after being confronted with their inconsistencies.

One of the most frustrating findings in the classical view of preferences is the *preference reversal* phenomenon, originally discovered by Lichtenstein & Slovic (1971) and Lindman (1971) and subsequently confirmed (see Slovic and Lichtenstein, 1983, for a review). In an example of the basic phenomenon, ($x, yY) denotes a delayed payment in which one receives $x in y years from now. Many people, when asked to choose between the delayed payments ($1,600, 1.5Y) and ($3,550, 10Y), prefer the first. However, when people are asked to state the instantaneous monetary value of these delayed payments, the majority assign a higher value to the second one (Tversky, Slovic, & Kahneman, 1990). Even when confronted with these seemingly contradictory actions, some people do not change their choices or evaluations but prefer to adhere to them and accept these paradoxical actions (Lichtenstein & Slovic, 1971).

Several explanations have been offered for these findings. At first, preference reversals were mostly explained as violations of transitivity. Recent support for this explanation was given by Loomes, Starmer, and Sugden (1989). A violation of *procedure invariance* is currently the prevailing explanation (Tversky et al., 1990). For instance, when assessing the instantaneous monetary values of delayed payments (*matching*), persons base their actions on a particular value system. When choosing between delayed payments, however, they base their actions on another value system. Therefore, preferences from choosing are different from those yielded by matching; this fact constitutes a violation of procedure invariance. The explanations invoking a violation of procedure invariance are mostly based on weighted additive models. These models assume that, both for choosing and for matching, a participant evaluates delayed payment ($x, yY) by using $av(x) + bw(y)$, where v and w are values for the separate attributes and $a > 0$ and $b > 0$ sum to 1 and are *importance weights*. It is generally assumed that the value functions v and w are the same for choosing and matching and that only the weights a and b vary. That is, for choosing, particular weights, such as a_c and b_c, are adopted, and for matching, other weights, such as a_m and b_m, are adopted. Note that because importance weights are positive, choosing and matching yield the same unidimensional orderings over single attributes (Tversky, Sattath, & Slovic, 1988, p. 377).

The *scale compatibility* hypothesis assumes that, in the matching task, the participant's attention is primarily directed toward the dimension for which a matching value is to be provided, that is, the monetary dimension in our example. The participant pays less attention to the other dimension (time), which therefore receives a lower importance weight. The scale compatibility hypothesis predicts that, in our example, with a_m being the importance weight for money in matching and with a_c being the importance weight for money in choosing, a_m will be higher than a_c. This prediction agrees with the observed choices.

Another explanation is provided by the *strategy compatibility hypothesis*. Choosing is basically a qualitative activ-

Peep F. M. Stalmeier, Nijmegen Institute for Cognition and Information (NICI), University of Nijmegen; Institute for Radiotherapy, St. Radboud Academic Hospital, Nijmegen, The Netherlands; and Medical Psychology, Amsterdam Academic Medical Centre, Amsterdam, The Netherlands. Peter P. Wakker, Medical Decision Making Unit, University of Leiden, Leiden, The Netherlands. Thom G. G. Bezembinder, NICI, University of Nijmegen, Nijmegen, The Netherlands.

This research was made possible through grant NUKC 93-136 from the Dutch Cancer Society. We thank Lia Verhoef, W. A. J. van Daal, and members of the Mathematical Psychology Department of the University of Nijmegen for helpful comments.

Correspondence concerning this article should be addressed to Peep F. M. Stalmeier, Nijmegen Institute for Cognition and Information, University of Nijmegen, P. O. Box 9104, 6500 HE Nijmegen, The Netherlands. Electronic mail may be sent via Internet to Stalmeier@nici.kun.nl.

ity; participants are more inclined to resort to qualitative methods for solving the decision problem. Hence, participants tend to simply choose the delayed payment that appears best for the prominent (most important) attribute, thus avoiding the necessity of comparing trade-offs for different attributes. For choosing, strategy compatibility leads to an overweighting of the prominent attribute; this overweighting is called the *prominence effect*. Matching is a quantitative activity; participants weigh a favorable trade-off for one attribute against an unfavorable trade-off for another attribute, so their value system is more balanced. On the basis of the assumption that time is the prominent attribute for delayed payments, the prominence effect again predicts that a_m will be higher than a_c, in agreement with the observed choices. In experiments by Fischer and Hawkins (1993), the prominence effect was found to be stronger than the scale compatibility effect.

Mellers, Ordóñez, and Birnbaum (1992) considered alternative explanations of preference reversals. In the *change of process theory*, fixed attribute valuation functions (v, w) (*scale convergence*) are assumed, in keeping with the weighted additive models described above. However, this theory permits more general ways of aggregating attribute valuation functions into overall valuations for various contexts. These investigators tested weighted average models (*contingent weighting theory*) for the general case in which the attribute valuation functions (v, w) can be any functions and rejected this theory in favor of the theory that the combination operation changes from multiplicative to additive depending on the response scale and stimulus distribution.

When a new phenomenon that falsifies an existing theory is discovered, as few modifications as possible are made and a new, generalized theory that is as simple as possible is developed. Thus, although preference reversals falsify classical preference theory, the explanations advanced so far have preserved most of preference theory. It is true that preferences are no longer considered invariable and that they are now considered to depend on the task (matching or choosing). However, within each task, preference theory is still accepted, and the only difference across tasks is in the variability of importance weights.

We describe preference reversals of a more extreme nature than in the cases described above. In the latter, choosing and matching tasks must coincide for comparisons in which only one attribute is involved. For instance, if a person prefers ($1,600, 1.5Y) to ($1,600, 5Y) in a choosing task, then the person should behave accordingly in a matching task. This prediction is a direct consequence of the higher scale value for 1.5Y than for 5Y (as also implied by scale convergence) and monotonicity for both contexts. This principle can be called *unidimensional procedure invariance*. The preference reversals presented in this article, however, violate unidimensional procedure invariance and thus scale convergence.

This article is organized as follows. First, we present the preference structure underlying our study and describe the experiments and results. We discuss the preference reversals revealed by the data and their implications for the foundations of decision theory. The relationship of our preference reversals to an interesting violation of dominance, described by Goldstein and Einhorn (1987) and by Birnbaum, Coffey, Mellers, and Weiss (1992), is explained. We describe implications of our discovery for commonly used methods of outcome evaluation in medical decision making and draw conclusions.

The Preference Structure Underlying Violations of Unidimensional Procedure Invariance

To obtain violations of unidimensional procedure invariance for choices between two-attribute alternatives, we used a domain in which the rational evaluation scheme is multiplicative rather than additive. Our stimuli concerned chronic states of health; for example, (L, Q) denotes living L years in the (constant) chronic state of health Q, followed by immediate death. A natural evaluation is the multiplicative scheme

$$V(L) \times W(Q), \tag{1}$$

where W(Q) evaluates well being in state of health Q and V(L) describes the valuation of life duration. The function V can deviate from linearity to express, for instance, time discounting (Miyamoto & Eraker, 1988).

Multiplicative models with positive factors can be reduced to additive models by a logarithmic transformation; therefore, they are equivalent to additive models in a preferential sense. In this study, however, states of health that are worse than death are considered and should be assigned negative values. In the presence of both positive and negative values, multiplicative models cannot be reduced to additive models.

It has been found that a rational valuation of chronic health states can be more complicated than the above-described multiplicative model. For a poor health state Q, there is usually a *maximum endurable time* S (depending on the participant), so that up to S the participant still values life positively (Sutherland, Llewelyn-Thomas, Boyd, & Till, 1982). In Figure 1, health states Q1 and Q2 are positive and the utility curves fan out; however, the utility curve for a health state with a maximum endurable time has a single peak. Let the value of (S, Q) be $U(S, Q) > 0$. Only for the years after S does the value of life become negative. Thus, for $T > S$ the total value of living T years in health state Q is

$$U(S,Q) + [V(T - S)] \times W(Q), \tag{2}$$

where W(Q) is negative. If the value of W(Q) is not highly negative and T is not much longer than S, then the described value of (T, Q) will still be positive and the number of healthy years equivalent to (T, Q) will be positive (shaded area in Figure 1). Only for the years after S can the valuation be determined by a negative multiplicative model.[1] Our

[1]Let L = T –S and subtract the constant U(S, Q); subtracting a constant can always be done for a utility function without affecting the preference scheme.

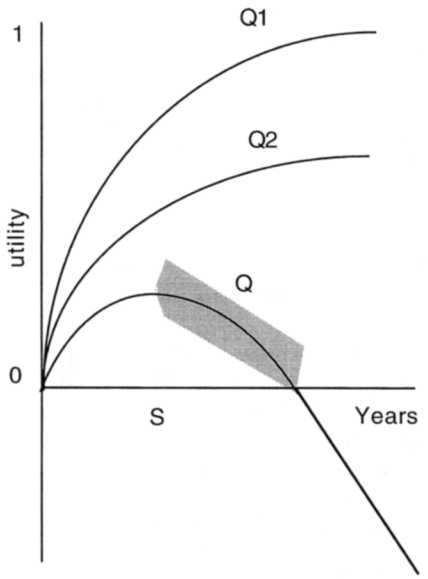

Figure 1. Utility versus duration for three different health states, Q1, Q2, and Q. For health state Q with a maximum endurable time S, the preference curve has a single peak.

unidimensional preference reversals are generated in the shaded area of Figure 1.

For this study, we selected participants for whom health durations were longer than S. For most of them, the value of stimuli (T, Q) still seemed to be positive; only a few participants did not have a positive number of healthy years equivalent to stimuli (T, Q).

Experiment 1a

In this experiment, women were selected by the condition that they preferred living 25 years with metastasized breast cancer, (25, M), to living 50 years with metastasized breast cancer, (50, M). Therefore, these women preferred a shorter life to a longer life for the health state metastasized breast cancer.

After being selected, women were asked how many years in good health they consider equivalent to the outcomes (25, M) and (50, M). Such questions are typically asked in the so-called time trade-off (TTO) method, which is commonly used in medical decision making to assess values of health states (Torrance, Thomas, & Sackett, 1972).

Method

Participants. Forty-eight women who were 20 years old or older participated. All participants were students at the University of Nijmegen.

Procedure. The interviews were done on an individual basis. In the selection session, which lasted for about 10 min, a participant was told that she would be participating in a pilot study concerning decision making by women who have an increased risk for breast cancer attributable to family history. Written health state descriptions containing the physical, psychological, and social consequences for three health states, namely, living with metastasized breast cancer, living after prophylactic mastectomy, and living with genetic counseling, were prepared. The participant read the health state description of living with metastasized breast cancer. She was selected for further participation if she preferred (25, M) to (50, M). In that case, the other two health state descriptions were given, and she was asked to read the health state descriptions carefully at home and imagine as vividly as possible how these health states would affect her personal life.

All health states were described as static, that is, the pain and discomfort were described as being constant over the whole duration. We also emphasized that death at the end of the 25- or 50-year period was not caused by the cancer process but that a painless death occurred through an unexpected accident. Thus, we countered the assumption that the first 25 years of life with cancer would be a better 25 years for a person who went on to live another 25 years than for a person who died after 25 years.

The test session was done on a separate day. The participant was asked to indicate the number of life years she would be prepared to sacrifice to avoid living for a given period in an inferior health state. The result is the TTO equivalent, (denoted by X), that is, the number of years in perfect health considered equivalent to Y years in inferior health. The number X was obtained by a *bisection procedure* involving *forced choices* between duration X in perfect health and fixed duration Y in inferior health (Sackett & Torrance, 1978). The starting number X was chosen randomly within the range from 0 to Y years to counterbalance anchoring effects. After an initial preference was stated, the value of X was bisected by the interviewer until the participant expressed indifference. The participants were instructed that there were no right or wrong answers and that their answers should reflect their own preferences. We used practice trials to ensure that participants understood the task.

The TTO test was administered with four durations Y for the three health states, namely, 5, 10, 25, and 50 years, amounting to 12 TTO questions. The questions were administered on a computer screen in a different random order for each participant. At the end of the test, the participants were again asked to indicate whether they preferred (25, M) or (50, M). The most preferred duration with metastasis (so, in terms of Equation 2, the maximum endurable time S) was determined by use of unidimensional choices among various durations of living with metastasis; e.g., "Do you prefer (5, M) or (10, M)?"

Results and Discussion

Thirty-four of 48 participants preferred the (25, M) outcome to the (50, M) outcome and were selected. The preferences of (25, M) over (50, M) were reliable in the sense that at the end of the experiment, all women but 1 confirmed the preference in the selection procedure for the shorter over the longer duration. Thus, 33 participants remained. We expected that these participants would assign a shorter TTO equivalent to the (50, M) outcome than to the preferred (25, M) outcome. Contrary to this expectation, the majority of the selected participants, 23 of the remaining 33, $p < .05$, assigned a longer TTO equivalent to the (50, M) outcome.

The (50, M) and (25, M) outcomes are unlikely because the 10-year survival rate for metastasized breast cancer is only 10%. In other words, our choice alternatives are hypothetical. For more realistic shorter life durations, similar preference reversals were observed; however, fewer participants passed the selection criterion. For instance, 11 par-

ticipants preferred (≤5, M) to (10, M). Seven of 11 assigned equal TTO equivalents, close to or equal to 0, to all four durations with metastasis. Contrary to the predictions of rational choice theories, of the remaining 4 participants, 3 assigned a longer TTO equivalent to the (10, M) outcome. Six participants preferred a duration of between 5 and 10 years with metastasis to (25, M). Contrary to the predictions of rational choice theories, 5 of these 6 participants assigned a longer TTO equivalent to the (25, M) outcome.

Seven participants were confronted with their preference reversals. Only 1 participant changed her responses, but not enough to alleviate the preference reversal.

Experiment 1b

In this experiment, we increased the number of participants to increase the power of our test. The experiment was conducted in a classroom setting with, on average, 17 students per class.

Method

Participants. Seventy-seven high-school students with a mean age of 18 years participated. Of these, 39 served as participants and the others played the role of interviewers.

Procedure. The students were divided into four groups instructed on separate occasions. A brief introduction to the medical decision-making problem for women at high risk for breast cancer was given. Two health state descriptions, prophylactic mastectomy and metastasis, were read by the students. Next, the TTO test with four durations, resulting in 8 questions, was administered on paper by the interviewers. The questions were presented to all participants in the same random order. There was ample time to ensure that the instructions were understood. After the test, the participants were asked to indicate whether they preferred (25, M) or (50, M).

Results

Of the 39 students, 17 preferred (25, M) to (50, M). The majority of these students (14 of 17, $p < .01$) assigned a longer TTO equivalent to the (50, M) outcome, a result that is contrary to the predictions of rational choice theories.

We carried out an analysis of the TTO X/Y ratios found in Experiment 1. This analysis provided another illustration of the irrationality contained in the participants' answers. The analysis also provided evidence for the *proportional heuristic* that may explain the participants' behavior in Experiment 1 and that we discuss later. The TTO equivalents from Experiment 1a and this experiment were combined in this analysis. We calculated the X/Y ratios for participants who, contrary to the predictions of rational choice theories, assigned a longer TTO equivalent to the (50, M) outcome. X denotes the TTO equivalent, and Y denotes the duration with metastasis. For instance, when a participant is indifferent for the choice between (5, Good Health) and (10, M), the X/Y ratio equals 5/10. The X/Y ratios are plotted in Figure 2. Note that, on average, the TTO equivalent X for (50, M) equals 27.2 healthy years, whereas

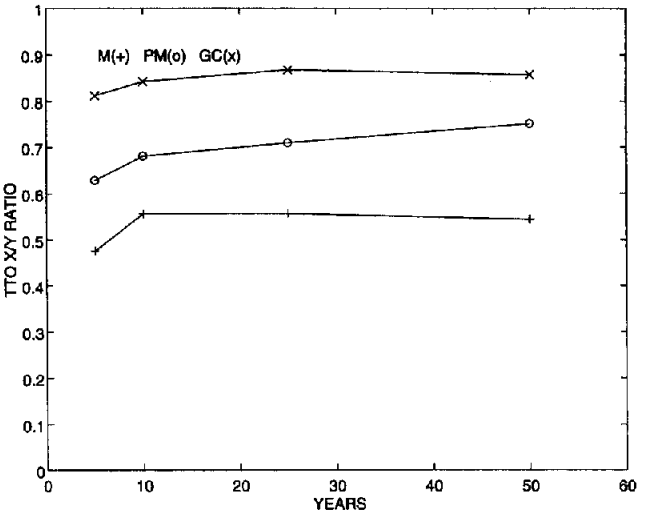

Figure 2. Time trade-off (TTO) X/Y ratios for participants who preferred (25, M) over (50, M). X = TTO equivalent; Y = duration with metastasis; M = metastasis; PM = prophylactic mastectomy; GC = genetic counseling.

the TTO equivalent X for (25, M) equals 13.9 healthy years, even though our selected participants preferred (25, M) over (50, M).

All three curves showed significantly lower X/Y ratios for the 5-year outcomes. Our young participants frequently commented that, for the short term, they disliked a decrease in the quality of life; for instance, they indicated that, for the short term, they would rather do without the burden of an operation with a long recovery period. Apparently, the value of the three health states should not be considered constant for the short term. Therefore, we used the 10-, 25-, and 50-year durations to test for constant X/Y ratios. We used multivariate within-subject tests for linear and quadratic contrasts. For the metastasis health state, the linear and quadratic contrasts were not significant, $F(1, 36) = 0.18$, $p < .67$, and $F(1, 36) = 0.19$, $p < .67$, respectively. For the prophylactic mastectomy health state, the linear contrast was significant, although the quadratic contrast was not, $F(1, 34) = 6.95$, $p < .01$, and $F(1, 34) = 0.16$, $p < .70$, respectively. For the genetic counseling health state, the linear and quadratic contrasts were not significant, $F(1, 37) = 0.63$, $p < .44$, and $F(1, 37) = 3.31$, $p < .08$, respectively.

Experiment 2

One might hypothesize that our participants tacitly assumed that living 25 or 50 years with metastasis is not realistic and that this assumption might have induced the preference reversal. Therefore, in this experiment, a more realistic health state was used. We chose living x days per week with migraine as that health state.

Method

Participants. Twenty-eight high-school students with a mean age of 18 years participated.

Procedure. This experiment was conducted in a classroom setting. The students read a 1-page health state description of living with migraine. Next, they were asked whether they preferred living 10 years with migraine x days per week to living 20 years with migraine x days per week. The number x was varied until 12 students preferred the shorter period. This preference occurred at 4 days per week with migraine. To slightly increase the number of students for whom the health state was negative, we subsequently used 4.5 days per week with migraine (4.5 Migr) in the TTO test.

A brief introduction to medical decision making for women who have an increased risk for breast cancer attributable to family history was given. The use of the TTO test was explained. The students served in turn as interviewers and as participants. The TTO test was administered on paper. Besides the migraine questions, additional questions were used with the genetic counseling and prophylactic mastectomy health states. Thus, the TTO test consisted of nine questions (three durations, 5, 10, and 20 years, and three health states, 4.5 days per week with migraine, prophylactic mastectomy, and genetic counseling). The questions were presented to all participants in the same random order. After the test, the participants were asked to indicate on paper whether they preferred living 10 years with migraine 4.5 days per week (10, 4.5 Migr) to (20, 4.5 Migr).

Results

Sixteen students preferred (10, 4.5 Migr) to (20, 4.5 Migr). Contrary to the predictions of rational choice theories, 15 of these 16 students, $p < .001$, assigned a longer TTO equivalent to the (20, 4.5 Migr) outcome than to the (10, 4.5 Migr) outcome. Only 1 student assigned a shorter TTO equivalent to the (20, 4.5 Migr) outcome.

Next, we analyzed the TTO X/Y ratios for this experiment. The TTO X/Y ratios for these 15 participants are plotted in Figure 3. Again, for the short term, our participants traded more healthy years. We used the 10- and 20-year durations to test for constant X/Y ratios. For the migraine health state, there was no significant difference, $F(1, 14) = 0.54$, $p < .48$. Also, for the prophylactic mastectomy health state, there was no significant difference, $F(1, 11) = 0.79$, $p < .39$. For the genetic counseling health state, the difference was significant, $F(1, 12) = 5.66$, $p < .04$.

Experiment 3

In this experiment, participants were confronted with the preference reversals and asked whether they wanted to change their response patterns.

Method

Participants. Sixty-four high-school students with a mean age of 17 years participated. Fifty-five of the participants were young women.

Procedure. This experiment was conducted in a classroom setting. The procedure was similar to that used in Experiment 2.

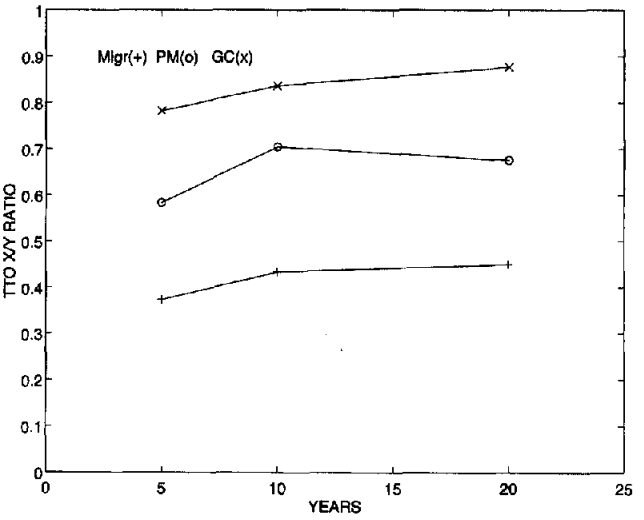

Figure 3. Time trade-off (TTO) X/Y ratios for participants who preferred (10, 4.5 Migr) over (20, 4.5 Migr). X = TTO equivalent; Y = duration with metastasis; Migr = migraine; PM = prophylactic mastectomy; GC = genetic counseling.

For the migraine health state, we chose 5 days per week with migraine.

The TTO test was administered on paper. Besides the migraine questions, additional questions were used with the genetic counseling and prophylactic mastectomy health states. The durations were the same as in Experiment 2. In contrast to the situation in our other experiments, the indifference point in the TTO test was obtained through a *direct match* and not through choices. This strategy was used to save time for the confrontation with the inconsistency. After the test, the participants were asked to indicate on paper whether they preferred (10, 5 Migr) to (20, 5 Migr).

Next, the nature of the preference reversal was explained to the participants. The following scheme was depicted on a blackboard:

$$\begin{array}{ccc} (10{,}5\text{ Migr}) & > & (20{,}5\text{ Migr}) \\ = & & = \\ (4,\text{ Healthy}) & < & (8,\text{ Healthy}). \end{array}$$

It was explained that the equals signs designated the indifference obtained in the TTO task between, e.g., (4, Healthy) and (10, 5 Migr). Next, it was explained that this response pattern entailed a preference reversal. This scheme was also present on the response sheet in the following form:

$$\begin{array}{ccc} (10{,}5\text{ Migr}) \ldots & & (20{,}5\text{ Migr}) \\ = & & = \\ (\ldots,\text{ Healthy}) \ldots & & (\ldots,\text{ Healthy}). \end{array}$$

The participants were instructed to indicate their choice preference for the first line (choosing task). They had to copy their corresponding answers from the matching task to the appropriate spaces in the last line of the scheme on the response sheet. Then, they filled in the appropriate preference signs. The participants then indicated whether their responses were inconsistent. In the case of inconsistency, they had to indicate on a 10-point scale whether they understood that their responses were inconsistent. In The Netherlands, the 10-point rating scale is the standard grading system, so our participants were familiar with this scale.

Finally, participants with inconsistent answers had to indicate

whether they were willing to change their response patterns. Four choices were offered: (a) no changes, (b) indicate a new match for (10, 5 Migr), (c) indicate a new match for (20, 5 Migr), and (d) indicate a new choice preference between (10, 5 Migr) and (20, 5 Migr).

Results

Sixty-four forms were returned. Three incomplete or illegible forms were discarded. Thus, 61 participants remained. Of these, 23 preferred (20, 5 Migr) to (10, 5 Migr) and 1 indicated that (10, 5 Migr) was equivalent to (20, 5 Migr).

For the remaining 37 participants, the shorter duration with migraine was preferred to the longer one, and a preference reversal was observed for 27 participants, $p < .01$. Of the remaining 10 participants, 5 assigned a longer TTO equivalent to the (10, 5 Migr) outcome and 5 assigned equal TTO equivalents (usually 0) to (10, 5 Migr) and (20, 5 Migr).

The 27 participants with the preference reversals all correctly indicated that they were inconsistent. The average rating for how well they understood that their responses were inconsistent was 7.5, which in the standard grading system in The Netherlands indicates a "more than sufficient" to "good" understanding. Of these 27 participants, 21 indicated that they did not want to change their responses. One participant changed her TTO equivalent for (10, 5 Migr) but remained inconsistent. Three participants changed their TTO equivalents for (20, 5 Migr); 2 of these 3 became consistent. Two participants changed both TTO equivalents and became consistent. Therefore, we concluded that after the confrontation, only 4 of 27 participants undid their original preference reversals. None of the 27 participants changed their original preference for (10, 5 Migr) over (20, 5 Migr).

General Discussion

For poor health states such as metastasized breast cancer or continuous migraines, we found that people may prefer a shorter to a longer life duration. Nevertheless, when asked to state the equivalent number of healthy years, participants as a rule demanded more healthy life years for the longer life duration than for the shorter one. Therefore, the longest number of healthy years was assigned to the nonpreferred outcome. The results of the three experiments taken together indicate that 79 of 103 participants assigned a longer life duration to the nonpreferred outcomes, $p < .001$. This finding is at odds with classical rationality principles and concurs with the other examples of preference reversals in the literature.

Possible Objections to Our Findings

Our results could be criticized because no real outcomes were used; therefore, participants might not have been highly motivated. Obviously, for the choice alternatives of our study, chronic health states, no "actual payment" is possible. Still, because of the practical relevance of this research, which was pointed out to the participants, our participants were motivated. The experiments in the literature have mostly considered hypothetical gambles for money. In general, no substantial difference was found between hypothetical gambles and gambles in which participants are actually paid according to their choices (Casey, 1991; Grether & Plott, 1979; Lichtenstein & Slovic, 1973; Mellers, Chang, Birnbaum, & Ordóñez, 1992; Pommerehne, Schneider, & Zweifel, 1982; Slovic, 1975, Experiment 3; Tversky & Kahneman, 1992); an exception is Bohm (1994).

Another concern with our experiments could be a selection bias. For example, suppose for a moment that the only "true" preference would be a preference for (20, 5 Migr) over (10, 5 Migr). Then, by selecting only participants who in a first preference expressed the opposite, we selected only incorrect prior preferences. A preference change by a majority of participants in the TTO test could be explained solely by a correction of preference on closer inspection by the participants. However, we carefully elicited the prior preference of (20, Migr) over (10, Migr). All participants but 1 reaffirmed their prior preference at the end of the experiment. Also, in Experiment 1b, the choice between (25, M) and (50, M) was obtained after the matching task. Therefore, we conclude that the described selection bias does not underlie our findings.

For the breast cancer experiment, one might object that, despite the static health state description, our participants perceived the metastasis health state as being less severe with the (50, M) stimulus than with the (25, M) stimulus. Some participants raised this objection and were told that the health state was really static. In any case, this objection does not explain why these participants still preferred the shorter to the longer life duration in the choosing task. Also, this objection cannot be raised for the migraine health state, for which similar rates of preference reversals were observed.

A factor that may have contributed partly to our findings is that the response scale only provides positive life durations, whereas some stimuli may actually have had a negative value (worse than immediate death) for some participants. In the preference model with a single-peak preference curve (Figure 1), the position of, for instance, (20, 4.5 Migr) would then be below the 0 level on the utility axis. Perhaps our experiment was misleading in this regard and so our findings are not relevant. However, we think that this aspect of a misleading positive response scale for negative stimuli may underlie some preference reversals, but not many. Only 10% of the participants matched (10, 4.5 Migr) to (0, Healthy). Most participants who exhibited preference reversals matched the migraine or metastasis health states to positive healthy life years considerably remote from 0; the average X/Y ratios were 0.4 and 0.55 for the migraine and metastasis health states, respectively. These results leave ample space on the response scale for matching (20, 4.5 Migr) or (50, M) to shorter numbers of healthy life years (shaded area in Figure 1). However, even after being con-

fronted with the inconsistency, most participants did not change their responses

Finally, it is possible that preference reversals are sensitive to context effects: Different sets of stimuli may induce different evaluation strategies (Birnbaum et al., 1992; Mellers, Ordóñez, & Birnbaum, 1992). We think that it is less likely that our violation is sensitive to context effects. Experiment 3 shows that the violation is robust against direct confrontation of the participants with the inconsistency; in a way, confrontation provides the most critical context for the phenomenon.

Implications of Our Findings

Our preference reversals were more extreme than traditional preference reversals. The (20, 5 Migr) and (10, 5 Migr) outcomes differed only in one attribute. Therefore, unidimensional procedure invariance was violated. In the choosing task, (10, 5 Migr) was preferred to (20, 5 Migr), but in the matching task, the participant assigned a higher value to (20, 5 Migr) than to (10, 5 Migr). Traditional preference reversals (Beattie & Baron, 1991; Birnbaum et al., 1992; Bostic, Herrnstein, & Luce, 1990; Casey, 1991; Cho, Luce, & von Winterfeldt, 1994; Delquié, 1993; Fischer & Hawkins, 1993; Fischhoff et al., 1980; Goldstein & Einhorn, 1987; Gonzáles-Vallejo & Wallsten, 1992; Grether & Plott, 1979; Hawkins, 1994; Hershey & Schoemaker, 1985; Johnson & Schkade, 1989; Lichtenstein & Slovic, 1971; Lichtenstein & Slovic, 1973; Lindman, 1971; Loomes et al., 1989; Mellers, Chang, et al., 1992; Mellers, Ordóñez, & Birnbaum, 1992; Pommerehne et al., 1982; Slovic, 1975; Slovic et al., 1990; Slovic & Lichtenstein, 1983; Tversky et al., 1988; Tversky et al., 1990) occur when stimuli are compared for more than one attribute, and they have been explained mostly by changes in the weighting with which the attributes are combined in different tasks. This explanation cannot be applied to our findings. (Note that importance weights must be positive so that sign reversals are excluded.)

The most central question for preference reversals is the following: If choosing and matching reveal different preferences, then which method provides "better" preferences? Here, "better" means more accurately reflective of the participant's values or more rational in a normative sense. Tversky et al. (1988, p. 383) asked respondents who exhibited preference reversals to reconsider their actions and found that modifications occurred in either direction. They conjectured that choosing and matching are both biased in opposite directions but suggested that their finding "may reflect a routine compromise rather than the result of a critical reassessment" on the part of the participants. In a study of choice compared with direct judgment of well being, Tversky and Griffin (1991) suggested that choice is "insufficiently sensitive to contrast" (p. 117) (contrast: compare your income to the average in your country), whereas judgment is insufficiently sensitive to endowment (endowment: absolute level of income) (see also Fischhoff, Slovic, & Lichtenstein, 1980; Johnson & Schkade, 1989, p. 423; and Slovic, Griffin, & Tversky, 1990, p. 25). In short, the present state of the art does not suggest whether choosing or matching is better for assessing preferences.

Our findings suggest that choosing provides the better information, at least in the domain of our study. Participants in our experiments were firm with regard to their initial preferences for the shorter duration over the longer duration for the metastasis and migraine health states (i.e., the selection criterion for the experiment), but when asked to explain the longer numbers of healthy life year equivalents, their reactions were different. Some replied, "One should take proportions here, shouldn't one?" Also, in Experiment 3, none of the participants changed their choice preferences after being confronted with the preference reversals, but several participants changed the number of healthy life year equivalents.

Our interpretation is as follows. In a matching task, a participant should provide a number as a response and usually should base this number on data that are also in the form of numbers. Apparently, this format induces participants to find simple mathematical heuristics for generating answers and to relate their actions insufficiently to their feelings and preferences. Even if the bisection procedure[2] is used, participants resort to mathematical heuristics, possibly recognizing the underlying matching task. In this study, when trading off life years, participants used a method that seems sensible in the majority of cases (positive health states) and did not realize that they were dealing with the exceptional case in which additional life years are not advantageous. This situation could be interpreted as an extreme case of scale compatibility, in which participants concentrate on the life duration dimension to the degree of forgetting that, because of the poor health state, additional life years should now be evaluated negatively. This interpretation suggests that, from a normative point of view, the heuristics adopted by participants in matching tasks are more ad hoc and irrational than has been commonly thought. This suggestion is a further argument in favor of the choosing task. In the section *Implications for Medical Decision Making*, we discuss in further detail a heuristic that the majority of participants seemed to adopt in our study, that is, the proportional heuristic.

Relationship With Birnbaum et al.'s (1992) Preference Reversal

Our preference reversal can be related to previous ones (Birnbaum et al., 1992; Goldstein & Einhorn, 1987, Experiment 3 [who refer to P. Slovic, personal communication, 1984]). These preference reversals are based on a violation

[2]Bostic et al. (1990) found that a similar method provides results closer to preferences than does a direct matching method, in which participants should immediately provide indifference values. Therefore, use of the bisection procedure critically tests our findings by maximally encouraging participants to act according to their preferences. If our matching responses are taken as genuine preferences, then our findings suggest violations of transitivity rather than of procedural invariance.

of monotonicity in choices between gambles. These preference reversals can also be interpreted as a violation of unidimensional procedure invariance.

The phenomenon was further tested and confirmed by Mellers, Weiss, and Birnbaum (1992), Birnbaum and Sutton (1992), and Birnbaum (1992). Birnbaum et al. (1992) studied configural weighting theory. A novel aspect of this theory is that 0 outcomes play a special role and can have decision weights different from those of other outcomes. If a 0 outcome in a lottery is replaced by a positive outcome, then, as a consequence, the certainty equivalent (CE) (i.e., the certain amount of money that is equivalent) of the lottery can decrease. This idea was confirmed by experiments and was most pronounced for the lottery (96, .95; 24) (receiving \$96 with probability .95, and \$24 with probability .05) against the lottery (96, .95; 0).

In a formal sense, it can be argued that the above-described finding implies a violation of unidimensional procedure invariance. As an explanation, suppose that we restrict attention to lotteries with a fixed probability, say .95, and suppress this probability. Then (X, Y) describes receipt of X with a probability of .95 and receipt of Y with a probability of .05. In the matching task, the typical finding is CE > CE', where CE and CE' are the certainty equivalents of (96, 0) and (96, 24), respectively. This finding, together with the obvious (96, 0) < (96, 24) in the choosing task, can be interpreted as a preference reversal that violates unidimensional procedure invariance.

Although our preference reversal as well as the above-described preference reversal can both be interpreted as a violation of procedure invariance, the underlying psychological processes are essentially different. Our finding is not based on neglect of a neutral outcome but rather on an entire sign reversal within a unidimensional scale. It is even more clear in the preference reversal of Birnbaum et al. (1992) than in ours that the irrational part is in the matching task of providing certainty equivalents. Again, in this quantitative task, participants resorted to mathematical heuristics and seemed to take some average of the outcomes in which, however, the 0 outcome was underweighted. Birnbaum et al.'s preference reversal was exhibited by up to 60% of participants. In a replication by von Winterfeldt, Chung, Luce, and Cho (1997), the percentage was somewhat lower; in several experiments it was about 35%. When the response mode did not involve direct quantitative judgment of certainty equivalents but involved binary choices, von Winterfeldt et al. found that the number of preference reversals was reduced. They suggested that the remaining ones could be explained by inconsistencies in choices. Note that we used binary choices in the matching task but still found preference reversals. Birnbaum and Sutton (1992) confronted participants with their preference reversals; most participants adjusted their actions so that the preference reversals disappeared. This finding is different from our finding, in which the preference reversals persisted for a large majority of participants when confronted with their inconsistencies. Thus, our preference reversals with poor health states may be more robust than the ones with gambles and neglect of the 0 outcome.

Implications for Medical Decision Making

The phenomenon described in this article was discovered in an empirical study of preference assessment methods for health states. Quality-adjusted life years are of considerable importance in the evaluation of health policies (see, e.g., Viscusi & Evans, 1990). They are based on subjective values of health states and life durations. One of the most popular methods of preference assessment is the TTO method, introduced by Torrance et al. (1972). Suppose a person states that, for life duration Y in state of health Q, X years in perfect health (X < Y) is equivalent to (Y, Q). Then, the quotient of X/Y is usually taken as a measure, W(Q), of the utility of health state Q, and (Y, Q) is valued by a general multiplicative form, $Y \times W(Q)$. The multiplication by W(Q) constitutes the quality adjustment of the life duration Y. Sometimes an adjusted TTO method is used; in this method, the utility of Q is taken as V(X)/V(Y) for a nonlinear function V that may, for instance, reflect time discounting, and (Y, Q) is measured by $V(Y) \times W(Q)$. These measurements are based on classical rational models for participants' answers to TTO questions. Our findings provide evidence against these classical models.

In Experiments 1 and 2, the answers of the 52 participants with preference reversals agreed well with a proportional evaluation in which, for a TTO question for (Y, Q), a participant chooses X as a proportion of Y, with the proportion W(Q) depending on Q. In the presence of response noise, for each fixed health state Q the answers X to TTO questions for (Y, Q) will be randomly distributed around a value, $Y \times W(Q)$, for W(Q) > 0. This proportional evaluation is supported by an ordinal analysis of the proportion X/Y. In Experiments 1a and 1b, 17 of 37 women assigned a lower proportion to the (50, M) outcome than to the (25, M) outcome, 5 women assigned equal proportions, and the remaining 15 assigned higher proportions. In Experiment 2, 6 participants assigned a lower proportion to the (20, 4.5 Migr) outcome than to the (10, 4.5 Migr) outcome, 3 participants assigned equal proportions, and the remaining 6 participants assigned higher proportions. These findings agree well with a proportional evaluation in which, because of random fluctuations in responses, a 50/50 split for the relative magnitude of the observed proportions is expected. The proportional heuristic is also supported by the analysis of the X/Y ratios. For the longer durations, we usually found a constant proportional trade-off. This finding is what one would expect for constant health states that are evaluated with a proportional heuristic. Note that a plot of X versus Y (not shown) results in a line pattern that fans out, indicating a multiplicative evaluation scheme (Anderson, 1976; Mellers, Ordóñez, & Birnbaum, 1992). In a number of studies, a constant proportional trade-off for various life durations in the inferior health state was accepted (Hall, Gerard, Salkeld, & Richardson, 1992; Pliskin, Shepard, & Weinstein, 1980; Verhoef, 1994). Significant but small deviations were found by Stalpers (1991), and stronger deviations were found by Sackett and Torrance (1978). Miyamoto and Eraker (1988) and McNeil, Weichselbaum, and Pauker (1981) found deviations for short life durations (5

years or less) in which participants did not want to give up any life duration in return for health improvement. We found lower X/Y ratios for the short-term duration of 5 years; our participants frequently commented that they were not willing to endure an illness for short durations.

The proportional evaluation hypothesis suggests that the ratio X/Y, used in the classical TTO approach, provides more reliable information about the participant than a "corrected" value V(X)/V(Y), because X/Y reflects the proportion chosen by the participant. Validity is a problem, however: In our experiments, the observed positive proportions clearly did not provide good measures for poor health states because the proportions of, for example, (10, 4.5 Migr) and (20, 4.5 Migr) were the same, whereas our participants preferred (10, 4.5 Migr). We conclude that for poor health states, the present usage of the TTO method is not appropriate.

Our preference reversals could be interpreted in terms of changes of scales or changes of processes to combine the scales (Birnbaum et al., 1992; Mellers, Ordóñez, & Birnbaum, 1992). To complete these theories, one needs to specify the scales and processes involved in both tasks. Because there is little structure in the present choice data, there is much freedom to speculate about whether scales or processes changed. Given the relatively clear violation of preference consistency, some explanation is needed, and given the current data, there are several possibilities that could be tested in future work.

Conclusion

This article has presented preference reversals of a more extreme nature than traditional preference reversals, because in our experiments unidimensional procedure invariance was systematically violated by the majority of participants. Therefore, they cannot be explained by models that leave the unidimensional scales intact, such as weighted additive models, and they require a more substantial modification of the classical preference model. Our data suggest that participants develop some rule of thumb for answering questions in an experiment and that the relationship between that rule of thumb and some hypothesized underlying preference system may be more remote and problematic than has been previously thought.

References

Anderson, N. H. (1976). How functional measurement can yield validated interval scales of mental quantities. *Journal of Applied Psychology, 61,* 677–692.

Beattie, J., & Baron, J. (1991). Investigating the effect of stimulus range on attribute weight. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 571–585.

Birnbaum, M. H. (1992). Violations of monotonicity and contextual effects in choice-based certainty equivalents. *Psychological Science, 3,* 310–314.

Birnbaum, M. H., Coffey, G., Mellers, B. A., & Weiss, R. (1992). Utility measurement: Configural-weight theory and the judge's point of view. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 331–346.

Birnbaum, M. H., & Sutton, S. E. (1992). Scale convergence and utility measurement. *Organizational Behavior and Human Decision Processes, 52,* 183–215.

Bohm, P. (1994). Time preference and preference reversal among experienced subjects: The effects of real payments. *Economic Journal, 104,* 1370–1378.

Bostic, R., Herrnstein, R. J., & Luce, R. D. (1990). The effect on the preference-reversal phenomenon of using choice indifferences. *Journal of Economic Behavior and Organization, 13,* 193–212.

Casey, J. T. (1991). Reversal of the preference reversal phenomenon. *Organizational Behavior and Human Decision Processes, 48,* 224–251.

Cho, Y., Luce, R. D., & von Winterfeldt, D. (1994). Tests of assumptions about the joint receipt of gambles in rank- and sign-dependent utility theory. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 931–943.

Delquié, P. (1993). Inconsistent trade-offs between attributes: New evidence in preference assessment biases. *Management Science, 39,* 1382–1395.

Fischer, G. W., & Hawkins, S. A. (1993). Strategy compatibility, scale compatibility, and the prominence effect. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 580–597.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1980). Knowing what you want: Measuring labile values. In T. Wallsten (Ed.), *Cognitive processes in choice and decision behavior* (pp. 119–141). Hillsdale, NJ: Erlbaum.

Goldstein, W. M., & Einhorn, H. J. (1987). Expression theory and the preference reversal phenomena. *Psychological Review, 94,* 236–254.

Gonzáles-Vallejo, C. C., & Wallsten, T. S. (1992). Effects of probability mode on preference reversal. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 855–864.

Grether, D., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review, 69,* 623–638.

Hall, J., Gerard, K., Salkeld, G., & Richardson, J. (1992). A cost utility analysis of mammography screening in Australia. *Social Science and Medicine, 34,* 993–1004.

Hawkins, S. A. (1994). Information processing strategies in riskless preference reversals: The prominence effect. *Organizational Behavior and Human Decision Processes, 59,* 1–26.

Hershey, J. C., & Schoemaker, P. H. J. (1985). Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science, 31,* 1213–1231.

Johnson, E., & Schkade, D. (1989). Bias in utility assessments: Further evidence and explanations. *Management Science, 35,* 406–424.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89,* 46–55.

Lichtenstein, S., & Slovic, P. (1973). Response-induced reversals of preferences in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology, 101,* 16–20.

Lindman, H. R. (1971). Inconsistent preferences among gambles. *Journal of Experimental Psychology, 89,* 390–397.

Loomes, G., Starmer, C., & Sugden, R. (1989). Preference reversal: Information-processing effect of rational non-transitive choice? *Economic Journal, 99(Suppl.),* 140–151.

McNeil, B. J., Weichselbaum, R., & Pauker, S. G. (1981). Speech

and survival: Trade-offs between quality and quantity of life in laryngeal cancer. *The New England Journal of Medicine, 305,* 982–987.

Mellers, B., Weiss, R., & Birnbaum, M. (1992). Violations of dominance in pricing judgments. *Journal of Risk and Uncertainty, 5,* 73–90.

Mellers, B. A., Chang, S., Birnbaum, M. H., & Ordóñez, L. (1992). Preferences, prices, and ratings in risky decision making. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 347–361.

Mellers, B. A., Ordóñez, L., & Birnbaum, M. H. (1992). A change-of-process theory for contextual effects and preference reversals in risky decision making. *Organizational Behavior and Human Decision Processes, 52,* 331–369.

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.

Pliskin, J. S., Shepard, D. S., & Weinstein, M. C. (1980). Utility functions for life years and health status. *Operations Research, 28,* 206–224.

Pommerehne, W., Schneider, F., & Zweifel, P. (1982). Economic theory of choice and the preference reversal phenomenon: A re-examination. *American Economic Review, 72,* 569–574.

Sackett, D. L., & Torrance, G. W. (1978). The utility of different health states as perceived by the general public. *Journal of Chronic Disease, 31,* 697–704.

Slovic, P. (1975). Choice between equally valued alternatives. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 280–287.

Slovic, P., Griffin, D., & Tversky, A. (1990). Compatibility effects in judgment and choice. In R. M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 5–27). Chicago: University of Chicago Press.

Slovic, P., & Lichtenstein, S. (1983). Preference reversal: A broader perspective. *American Economic Review, 72,* 923–955.

Stalpers, L. J. A. (1991). *Clinical decision making in oncology.* Unpublished doctoral dissertation, University of Nijmegen, Nijmegen, The Netherlands.

Sutherland, H. J., Llewelyn-Thomas, H., Boyd, N. F., & Till, J. E. (1982). Attitudes toward quality of survival: The concept of "maximum endurable time." *Medical Decision Making, 2,* 299–309.

Torrance, G. W., Thomas, W. H., & Sackett, D. L. (1972). A utility maximization model for evaluation of health care programs. *Health Services Research, 7,* 118–133.

Tversky, A., & Griffin, D. (1991). Endowment and contrast in judgments of well-being. In F. Strack, M. Argyle, & N. Schwarz (Eds.), *Subjective well-being* (pp. 101–118). Elmsford, NY: Pergamon Press.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 7,* 297–323.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95,* 371–384.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *American Economic Review, 80,* 204–217.

Verhoef, L. C. G. *The measurement of individual preferences for treatment outcomes in breast cancer.* Unpublished doctoral dissertation, University of Nijmegen, Nijmegen, The Netherlands.

Viscusi, W. K., & Evans, W. N. (1990). Utility functions that depend on health status: Estimates and economic implications. *American Economic Review, 80,* 353–374.

von Winterfeldt, D., Chung, N. K., Luce, R. D., & Cho, Y. (1997). Tests of consequence monotonicity in decision making under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 406–426.

Received July 12, 1995
Revision received February 12, 1996
Accepted May 28, 1996 ■