TI 2004-038/3
Tinbergen Institute Discussion Paper

Inconsistent and Lexicographic Choices in Stated Preference Analysis

*Jan Rouwendal**
Arianne T. de Blaeij

Department of Economics, Vrije Universiteit Amsterdam.

** Tinbergen Institute.*

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Please send questions and/or remarks of non-scientific nature to driessen@tinbergen.nl.
Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

Inconsistent and lexicographic choices in stated preference analysis

Jan Rouwendal* and Arianne T. de Blaeij

Department of Spatial Economics
Free University
Amsterdam

This version: March 30, 2004.

Keywords: stated preference analysis, choice experiments, inconsistent choices, lexicographic choices, value of time, value of a statistical life
JEL Classification Codes: C25, C93, Q51, R41

Abstract
In stated choice (SC) data inconsistent and lexicographic choice behavior is often observed. It is sometimes recommended to remove data with these characteristics from the analysis. In this paper we reconsider this recommendation. In our data many respondents have inconsistent choice patterns, which appear to be due to incidental mistakes. Moreover, a large number of the consistent respondents have lexicographic choice patterns. We show that the logit model, which is the most popular tool for analyzing SP data, is compatible with inconsistent and seemingly lexicographic choice behavior and that it offers precise predictions about the occurrence of such choices. In the data at our disposal the actual number of respondents who made different choices in two identical choice situations is substantially *lower* than that predicted by the model, whereas the number of respondents with lexicographic answers is much larger than predicted. The logit model is then adapted in various ways to bring it in better agreement with the facts. In particular, we introduce an effect of remembering the earlier choice when the same situation recurs, the presence of latent classes of lexicographic respondents and the presence of heterogeneity among respondents.

Corresponding author: Jan Rouwendal, Department of Economics, Free University, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, tel 31-20-4446093, fax 31-20-4446004,email jrouwendal@feweb.vu.nl.

* Jan Rouwendal is also at Wageningen University.

1 Introduction

Investigation of stated preferences has become a popular tool for economic analysis of evaluation problems in environmental economics, transportation, marketing and other fields. One stated preference methodology is the stated choice (SC) method. The stated choice method asks the respondents to make a choice between bundles of attributes. Usually a respondent is asked to make a number of such choices. The stated choice methodology is therefore one of the attribute based methods for preference valuation. The focus of such methods is to willingness to pay or related welfare economic concepts (see, for instance, Holmes and Adamowicz, 2003). A crucial assumption for attribute based method is is that a respondent's willingness to pay is related to his of her underlying preferences in a consistent manner. A detailed discussion of stated choice methods is provided in Louviere et al. (2000).

In environmental economics SC is used to value environmental goods and services. The value of many aspects of the natural area is not reflected in market prices. To value these aspects, the valuation method measures the respondent's willingness to pay for an increase in the quality of quantity, or a related concept such as the willingness to accept a decrease in quality or quantity. Similar procedures are used in marketing, transportation and other fields. By a careful design of the SC experiment, a researcher is able to make inferences into aspects of behavior that cannot yet be observed, such as the consumer's reaction to the introduction of new products.

Another important advantage of SC is that the data can often be collected in a relatively inexpensive way and under circumstances that are in some respects close to that of a laboratory. Moreover, it is usually possible to ask the same respondent about choice behavior in a number of different choice situations, whereas most data referring to actual behavior inform a researcher only about behavior of subjects in one specific set of circumstances. The data collected in SC experiments can therefore in some respects be 'richer' than data based on actual behavior and this is an important reason why SC analysis is a welcome addition to the toolkit of economists.

However, it is also clear that stated choice experiments have their own weaknesses. Perhaps the most important of them is that respondents do not have exactly the same incentives as in the real life situations they are supposed to imagine. There is a danger that respondents will answer on the basis of the perceived preferences of the researcher (who sometimes pays them for participation) or of their reference groups, while the consequences a choice would have for themselves get a smaller weight than would be the case in real life situations. The respondents may also get tired or bored by having to answer many questions about an issue in which they are not really interested. As a result they may not make deliberate choices, but answer after taking only a superficial look at the problem posed. For these reasons the outcomes of stated preference experiments could be less reliable than information about actual behavior.

Some work has been done to distinguish the unreliable decision-makers from the rest. Researchers have looked to lexicographic and inconsistent choices with this purpose in mind. Lexicographic choices occur when the respondent always chooses the alternative that is best (or worse) with respect to one characteristic. This may be the result of using a rule of thumb that allows one to proceed fast through the questionnaire, but it may also indicate true lexicographic preferences. In the latter case the respondent is of the opinion that usually one aspect (for instance, safety in a questionnaire referring to travel behavior) should dominate her choices and that no trade-off between this aspect and others is possible.

Saelensminde (2002) defines inconsistent choices as choices that violate the transitivity axiom of consumer theory. Inconsistent choices may be detected in various ways. Sometimes researchers have deliberately included the same choice situation twice in the questionnaire in order to be able to check consistency in a straightforward way; others have included a dominant choice alternative

(see Rizzi and Ortuzar, 2003). A more sophisticated test combines the information contained in every choice a respondent makes and checks for contradictions (Saelensminde, 2001). Also in this case it is not exactly clear which interpretation must be given to the observation of inconsistencies. Respondents who attempt to make deliberate choices may incidentally make mistakes, and this may render the set of their choices inconsistent. On the other hand, it may happen that uninterested respondents answer the two identical questions identically by pure chance.

It is sometimes argued in the literature that more reliable research results will be obtained if the researcher removes the lexicographic or inconsistent observations from the data set. It has been shown repeatedly that estimates of the research targets, such as the monetary value of environmental damage, are influenced by this procedure (see, for example, Johnson et al, 2000; Rizzi and Ortuzar, 2003; Saelensminde, 2001). These results are obtained on the basis of the estimation of logit models. However, it will be pointed out below that standard applications of this model incorporate an implicit assumption about errors involved in the choice process that would lead one to expect that lexicographic and inconsistent choices occur (with precisely predicted frequencies). This means that if the data are generated in the way presumed by the model, removing the lexicographic and inconsistent choices may well *result in* biased estimates, instead of preventing them.

In this paper we take a different look at the problem. We start by asking the question whether it is meaningful to hypothesize that respondents make fully consistent choices when answering a SC questionnaire. An empirical answer is provided by applying a consistency check on the data at our disposal, in which respondents have 10 times made a choice between two possible routes that differed in three aspects. The choice task involved is certainly not extreme for SC questionnaires. The results show that a majority of respondents that did not make lexicographic choices showed at least one inconsistency. This means that removing all inconsistent respondents would leave us with a sample in which many respondents are lexicographic, which is unsatisfactory. On the other hand, the inconsistent respondents are often close to consistent in the sense that one different choice would have made their choice patterns consistent.

We conclude from this exercise that it is not useful (at least not for the data at hand) to proceed on the basis of the assumption that the choices made by the respondents are generated without error by a preference ordering. Instead we adopt the error generating mechanism that leads to the logit model. This model is the standard tool for analysis of SC data. We derive the predictions about occurrence of lexicographic choices and providing different answers to identical questions implied by this model and compare them to the observed frequencies of these phenomena. We illustrate the basic methodology with a simple version of the logit model and find that there are substantially more lexicographic respondents in our sample than predicted by the model. However, the actual number of respondents who made different choices in identical situations is much *smaller* than predicted.

In the remainder of the paper we attempt to find a reliable explanation for the lower-than-predicted number of respondents who made different choices in two identical choice situations. We consider the possibility that many respondents remember what they choose the first time and make the same choice again and incorporate it in our model. Moreover, we allow for the presence of latent classes of respondents with lexicographic choice behavior in our population. Introduction of these additions leads to a considerable improvement in model performance. We also investigate the role of observed and unobserved heterogeneity among the respondents as a possible explanation for the different responses in identical situations and for lexicographic behavior.

We proceed as follows. The next section provides a general discussion of the analysis of SC data under the assumption of a deterministic and a probabilistic choice mechanism. The former is shown to enable a check on the overall consistency of choice sequences of respondents that is applied in section 4. The latter leads to the logit model that is used in subsequent sections. Section 3 contains a discussion of the data we use in this paper.
In section 5 we compare the predictions of a simple logit model with respect to different choices in identical circumstances and lexicographic behavior with the actual frequencies. In section 6 we make similar comparisons for versions of the logit model that take into account observed and unobserved heterogeneity. Section 7 concludes.

2 Interpretation of SP data

Deterministic choice mechanism

The basic hypothesis is that choices made by respondents reflect a preference ordering over all alternatives. We assume that a respondent has preferences over all choice alternatives that can be described by means of a utility function *u*, with the attributes *x* of the alternatives as its arguments. The preferences may be dependent on the respondent's (observed and unobserved) characteristics *z*. Alternatives are distinguished by means of a suffix and the utility of alternative *i* is:

$$u_i = u(x_i; z) \tag{1}$$

If the consumer has to choose between two alternatives, *i* and *j*, and does so on the basis of his preferences, he chooses the alternative for which utility is highest.
Let $I(i|i,j)$ be an indicator variable that takes on the value 1 when alternative *i* is chosen and is equal to 0 otherwise. According to the theory just laid out, we should have:

$$I(i \mid i, j) = 1 \quad \text{iff} \quad (u_i - u_j) > 0 \tag{2}$$

The bulk of the literature on SP data assumes that preferences are linear in the attributes, and we will adopt this assumption throughout the paper. It means that we can write the utility function of the respondent as:

$$u(x_i, z) = \sum_{n=1}^{N} \beta_n(z) x_{ni} \tag{3}$$

where N is the number of attributes and the βs are coefficients that may be respondent-specific. It follows from (2) that:

$$I(i \mid i, j) = 1 \ \text{ iff } \ \sum_{n=1}^{N} \beta_n(z)\left(x_{ni} - x_{nj}\right) > 0 \tag{4}$$

If the respondent makes *M* choices, her choices define *M* inequalities that should be consistent with each other. This imposes testable restrictions on the data. This seems to have been pointed out first by Bates (1994) who constructed 'ray diagrams' for the case in which there are three characteristics.[1] Each choice defines a ray and this ray is the boundary of the combinations of the parameters β of the respondent's utility function that are consistent with her choice.

In general, the equation $\sum_{n=1}^{N} \beta_n(z)\left(x_{ni} - x_{nj}\right) = 0$ defines a hyperplane in the space to which the coefficients β belong. The βs that are consistent with a choice made by the respondents are in one of the two open half-spaces defined by this hyperplane. Each choice of the respondent defines

[1] The test of rationality this suggests is similar in spirit to that developed by Varian (1982) who considered utility maximization under a budget restriction.

such a half-space. Consistency requires that the intersection of the half-spaces defined by all questions is not empty. An elementary violation of this requirement occurs when a respondent makes different choices in identical choice situations: the half-spaces defined by these choices are disjoint and their intersection is therefore empty. It should be noted that this consistency check allows for individual-specific coefficients β.

Sælensminde (2001, 2002) used a partial version of this check by comparing only two choices at a time.[2] Sælensminde (2001) found that more than 60% of the respondents who had to make nine pairwise comparisons made at least one choice that is inconsistent with another choice, whereas Sælensminde (2002) found that more than one third of respondents who had to make four pairwise comparisons made at least one inconsistent choice.

Foster and Maurato (2002) consider some other tests for the logical consistency of SP choices. If one alternative dominates the other in the sense that all characteristics have a value that is equal to or better than that of the other, it should always be chosen. If alternative A is preferred to B, and another pairwise comparison results in B being preferred to C, than a pairwise comparison of A and C should result in a choice for A.[3] These dominance and transitivity requirements are aspects of the general consistency test based on (4) that was discussed above. It was found that 17% of the respondents violated the dominance criterion and 13% the transitivity requirement. Given the elementary nature of especially the dominance criterion, these fractions are surprisingly high.

Lexicographic choices are always consistent. They are compatible with a linear utility function (3) if one of the parameters, say the first, is such that $\beta_1(x_{1i} - x_{1j})$ always exceeds $\sum_{n=2}^{N} \beta_n(z)\left(x_{ni} - x_{nj}\right)$ in absolute value. This means that the occurrence of lexicographic choices is also affected by the values chosen for the attributes. If the real preferences of a respondent are given by (3), the trade-offs she is willing to make can be uncovered by using the 'right' values of the attributes. However, if a respondent has truly lexicographic preferences, such trade-offs do not exist. It is therefore important to check whether lexicographic choice behavior is compatible with reasonable values of the trade-offs involved.[4]

Probabilistic choice mechanism

The fact that even partial and elementary checks show many violations of consistency suggest that it is worthwhile to consider explicitly the possibility that the choice process is subject to errors. Fortunately, the logit model, which is the workhorse of stated preference analysis, can be interpreted as incorporating such a mechanism. In its standard form this model can be related to the discussion above by adding an error term ε to the behavioral hypothesis (2):

$$I(i \mid i, j) = 1 \quad \text{iff} \quad (u_i - u_j) + \sigma\varepsilon > 0 \qquad (5)$$

and assuming that ε is logistic distributed. σ is a scaling factor.[5] When σ=0 this reduces to (2). For positive values of σ this model implies choice probabilities of the form:

---

[2] He restricts attention to the non-emptiness of the intersection of two half-spaces, whereas a complete consistency check would investigate the non-emptiness of the intersection of all the half-spaces defined by the respondent's choices.

[3] A third test considered by Foster and Maurato (2002) refers to the ranking of more than two alternatives. This test is not considered here.

[4] See Rosenberger et al. (2003) for a discussion of lexicographic choice behavior.

[5] The introduction of the random term may be interpreted as the result of an imperfect ability to choose, see De Palma et al. (1994).

$$\Pr(I(i \mid i,j)=1)=\frac{e^{(u_i-u_j)/\sigma}}{1+e^{(u_i-u_j)/\sigma}} \qquad (6)$$

In applications of the model a standard assumption is that the random terms of each choice are independent. The value of the scaling factor reflects the variance of the error term. It seems natural to relate it to the complexity of the choice process and DeShazo and Fermo (2002) have indeed shown that its value increases when choice situations become more complex.

An important implication of (6) is that it is compatible with, and indeed predicts the occurrence of different choices in identical choice situations. This is clear from the fact that $u_i - u_j$ can be positive in (6), while the probability that alternative *i* will be chosen is still smaller than 1. In general it can only be said on the basis of (6) that $u_i - u_j > 0$ implies that the probability that *i* will be chosen exceeds the probability that *j* will be chosen. This much weaker relation between the preferences of the respondent and her choice behavior is a consequence of the 'noise' caused by the random term ε.

Let *p* be the probability that a respondent chooses alternative *i*, when confronted with alternatives *i* and *j*. If she faces the same choice situation twice, there are three possibilities:

- alternative *i* is chosen twice, the probability of this event is p^2
- alternative *j* is chosen twice, the probability of this event is $(1\text{-}p)^2$
- different choices are made, the probability of this event is $2p(1\text{-}p)$.

The respondent may therefore be consistently right (if she chooses the alternative with the highest probability twice), consistently wrong (if the alternative with the lowest probability is chosen twice) or inconsistent. According to the model, the probability of choosing consistently right is higher than that of choosing consistently wrong. The probability of making two different choices is at least twice as high than that of choosing consistently wrong and exceeds that of choosing consistently right if the probability of choosing right is lower than 2/3. These are strong predictions that can be checked easily if respondents are placed twice in the same choice situation without being aware of it.

The introduction of a probabilistic choice mechanism has as a consequence that the predicted frequencies of lexicographic choice behavior become small, even for a moderate number of choice situations. This means that unless respondents have indeed preferences that are truly lexicographic or close to lexicographic in that they make extreme trade-offs between the various attributes of the choice situation, it becomes unlikely that such choice behavior will ever be observed.[6] For instance, even if a particular respondent has a probability .95 of making the 'lexicographic' choice in each situation, the probability that his sequence of choices is lexicographic when 10 situations are presented to him is .59 only. If the probability for each particular choice is .9, the probability for the sequence is only .39.

3 Data

The data we use in this paper were gathered as part of a larger survey carried out by a specialized bureau (Intomart). This company has organized a large panel of respondents who are paid for filling out regularly a questionnaire. The information used here refers to a number of stated choice questions that were formulated in order to investigate the respondent's valuation of changes in traffic risk. Each respondent was asked to imagine that she has to make a trip from A

[6] This assumes that respondents choose in accordance with their preferences. Unreliable respondents who make lexicographic choices because they use a rule of thumb that allows them to proceed fastly through the questionnaire are indistinguishable from truly lexicographic respondents.

to B by car, while there are no other persons in the car. She can use two roads, which may differ in three attributes: toll, risk of a fatal accident and travel time. The respondent is told that these three attributes are the only ones in which the two roads differ. Travel time was included in order to facilitate a comparison of the outcomes of this study with travel time valuation studies. Since there are no other Dutch studies on the value of a statistical life, such a comparison was thought to be desirable.

The purpose of the survey was to measure the value of a statistical life and for this reason the differences in the toll and the number of fatalities were always of the opposite sign.[7] When there was also a change in travel time, it had the same sign as the change in the number of fatalities. This means that the choice situations posed to the respondents implied that they always had to pay for additional safety and less travel time.

Each respondent was ten times asked to make a choice between two roads. Five sequences of 10 choice situations were formulated. In each sequence the second and the tenth situations were identical. Each respondent was randomly assigned to one sequence. Thjis defines five groups of respondents. The size of each group is indicated in Table 1.

The three attributes were specified as follows. The toll is the price per trip in Dutch guilders.[8] It varies between Dfl 2.50 and Dfl 12.50. The travel time varies between 50 minutes and 1 hour. Road safety is indicated by the annual number of fatalities, which varies between 12 and 36. The respondents were informed that the number of trips made on the road during a year is 18 million, which means that the lowest number of fatalities (12) corresponds to the average safety level on Dutch roads.

It may of course be doubted whether the respondents were able to do exactly what the researchers asked them. Two roads that differ in only the three attributes mentioned are perhaps hard to imagine. For instance, it is likely that differences in the number of fatal accidents are correlated in reality with differences in the number of non-fatal accidents. Dutch drivers are not used to toll paying, except for a few bridges and tunnels with small traffic flows.

There were 1055 respondents who all answered the ten questions. Non-response was absent since respondents had to answer in order to be able to proceed to the end of the questionnaire (and consequently receive their payment). The necessity to give a response may have had a deteriorating effect on the quality of the responses and increases the desirability of carrying out a reliability check.

Table 1 Basic information about the data

Group	Number of respondents	Different choices in 2 and 10	Always lowest toll	Always lowest # fatalities	Always lowest travel time
1	207	33	16	27	53
2	220	36	15	44	46
3	211	20	22	35	37
4	215	41	19	39	45
5	202	29	31	36	42
Total	1055	159	103	181	223

Note. The number of choice situations for which there is no difference in travel time between the two alternatives for groups 1-5 is 3,3,1,4 and 5, respectively.

[7] This implies that dominating alternatives do not occur.

[8] 1 Dutch guilder was .45 euro.

Table 1 also provides information with respect to the number of respondents that made different choices in situations 2 and 10 and on the numbers that always choose the alternatives that scored best on one of the three attributes. Approximately 15% of the respondents made different choices in situations 2 and 10. Almost 10% of the respondents always choose the alternative with the lowest toll and 21% always choose the alternative with the lowest travel time. The alternative with the lowest travel time is always the one with the lowest number of fatalities, so the latter number also contains all respondents that have always opted for the alternative with the lowest number of fatalities. It should be remembered that such one-sided choice behavior is not necessarily generated by pure lexicographic preferences, that is preferences in which there is no trade-off between the scores of the various attributes. With a sufficiently large value for one parameter relative to the others a linear utility function (as defined in (3)) can lead to choices that are seemingly lexicographic.

Throughout the paper we will maintain the hypothesis that, unless preferences are pure lexicographic, the utility of each of the chosen alternatives is a linear function of the three attributes:

$$u(x;z) = \beta_1(z)toll + \beta_2(z)\,fatalities + \beta_3 travel\ time \qquad (7)$$

The value of a statistical life (*vosl*) and the value of travel time (*vot*) implied by this utility function are:

$$vosl = \frac{\beta_2(z)}{\beta_1(z)}, \qquad vot = \frac{\beta_3(z)}{\beta_1(z)} \qquad (8)$$

4 A first consistency check

We start our analysis of the data by carrying out a consistency check based on the linear inequalities defined in (4). This means that we specify the difference $u_i - u_j$ as:

$$u_i - u_j == \beta_1(z)\Delta toll + \beta_2(z)\Delta fatalities + \beta_3(z)\Delta time \qquad (9)$$

The choices made by the respondent define the following inequalities:[9]

$$\text{if } I(i \mid i,j) = 1: \quad vosl > -\frac{\Delta toll}{\Delta fatalities} - vot\frac{\Delta time}{\Delta fatalities} \qquad (10a)$$

$$\text{if } I(i \mid i,j) = 0: \quad vosl < -\frac{\Delta toll}{\Delta fatalities} - vot\frac{\Delta time}{\Delta fatalities} \qquad (10a)$$

These linear inequalities have a positive intercept since the changes in the toll and the number of fatalities have opposite signs. If there is a difference between the travel times of the roads among which the respondents have to choose, it has the same sign as the change in the number of fatalities, which implies that the slope is negative.

The inequalities (10) tell us that each choice made by the respondent contains information about her combination of *vosl* and *vot*. If her preferences are given by the linear utility function (3) and no mistakes are made in choosing, the ten inequalities should all be consistent.

If a respondent has chosen different alternatives in questions 2 and 10, this implies two contradictory inequalities for her combination of *vosl* and *vot* and her preferences are clearly inconsistent. However, this is clearly not the only case in which inconsistencies can occur.

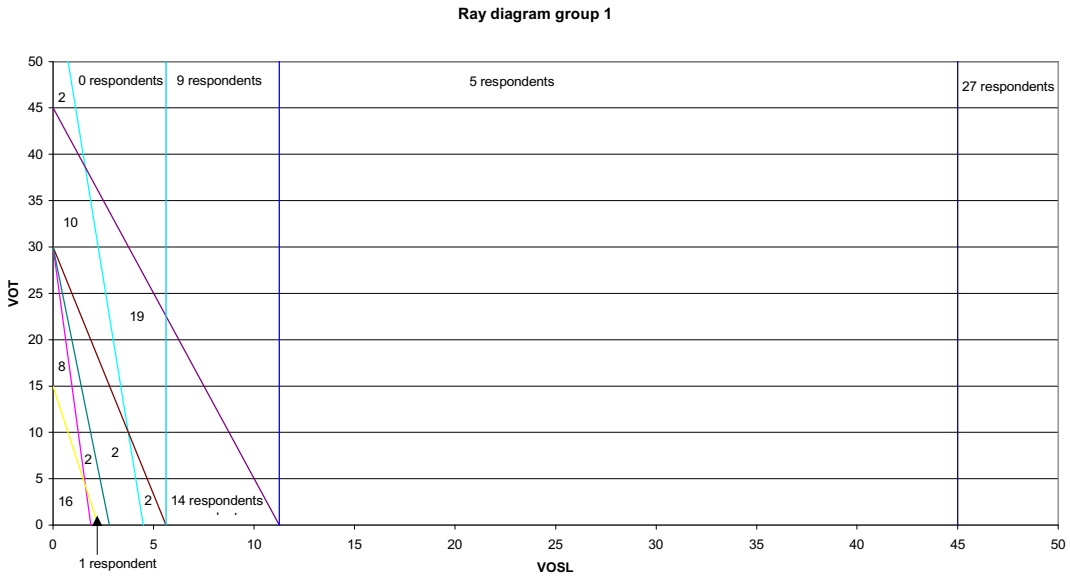
[9] If *Δfatalities* is positive. If it is negative, the inequality signs must be reversed.

It should be noted that this consistency check assumes a linear utility function, but allows the parameters of this function to be individual-specific without imposing any restrictions on the distributions of these coefficients.

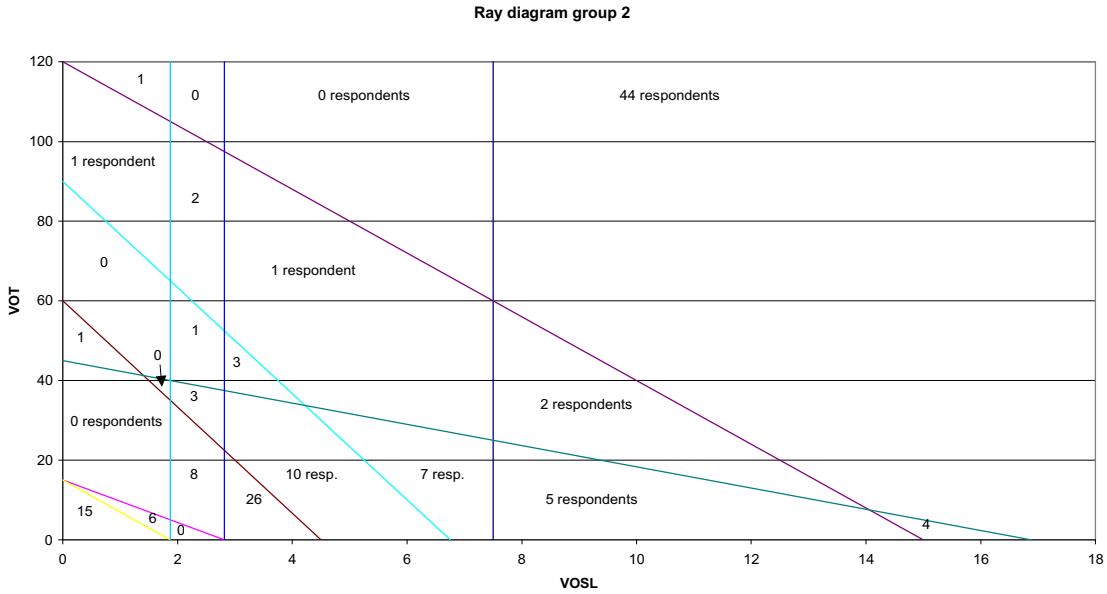
Table 2 Inconsistent and lexicographic respondents

Group	*n*	Overall Inconsistent	Overall Consistent	Lexicographic
1	207	90	117	43
2	220	80	140	69
3	211	93	118	57
4	215	79	136	58
5	202	66	136	67
Total	1055	391	664	294

Table 2 presents the results of the consistency check. It shows more than one third of the respondents have made at least one choice that is inconsistent with the other choices. In comparison with the results reported in the literature discussed above, this is not a bad result. More than one third of consistent choices are lexicographic choices. Figure 1 presents the ray diagrams and the consistent choice for each of the five groups are presented. The rays are defined by equations 8. Each choice implies that the combination of *vosl* and *vot* of the respondent lies above or below such a ray. For the lexicographic respondents combinations above or below all rays are relevant. It appears from this Figure that the number of lexicographic choices is much larger than one would expect if the simultaneous distribution of *vot* and *vosl* were smooth and unimodal as, for instance, a bivariate normal distribution. This may be interpreted as suggesting that the respondents with lexicographic choice behavior belong to separate subgroups in our sample. We will return to this suggestion below, when we estimate logit models.
How serious are the violations of the consistency requirement in our sample? In answering this question it may, first of all, be noticed that there are 2^{10}=1024 ways to make 10 times a choice out of two alternatives. Each possibility of making such choices corresponds with an array of zeros and ones consisting of ten elements and we refer to such an array as a choice pattern. The number of consistent choice patterns, that is choice patterns that correspond with consistent choices, differs over the five groups and equals 14, 23, 19, 30 and 22 for group 1,2,..,5, respectively. The probability that a consistent choice pattern will be chosen purely by chance is therefore small and in all cases lower than 3%.
For any respondent with an inconsistent choice patterns we can determine the seriousness of the violation of the consistency requirement by computing the minimum number of choices that would have to be different in order to make the choice pattern consistent. The results of this computation are shown in Table 3. More than tow thirds of the inconsistent respondents in all groups would have been consistent if one choice had been different. The maximum number of choices that would have to be different in order to make every respondent consistent is 3. Although this does not prove (in any formal sense of the word) that the respondents with inconsistent choice patterns are in fact consistent decision makers who incidentally make mistakes, in our view the figures shown in the Table strongly suggest this interpretation.

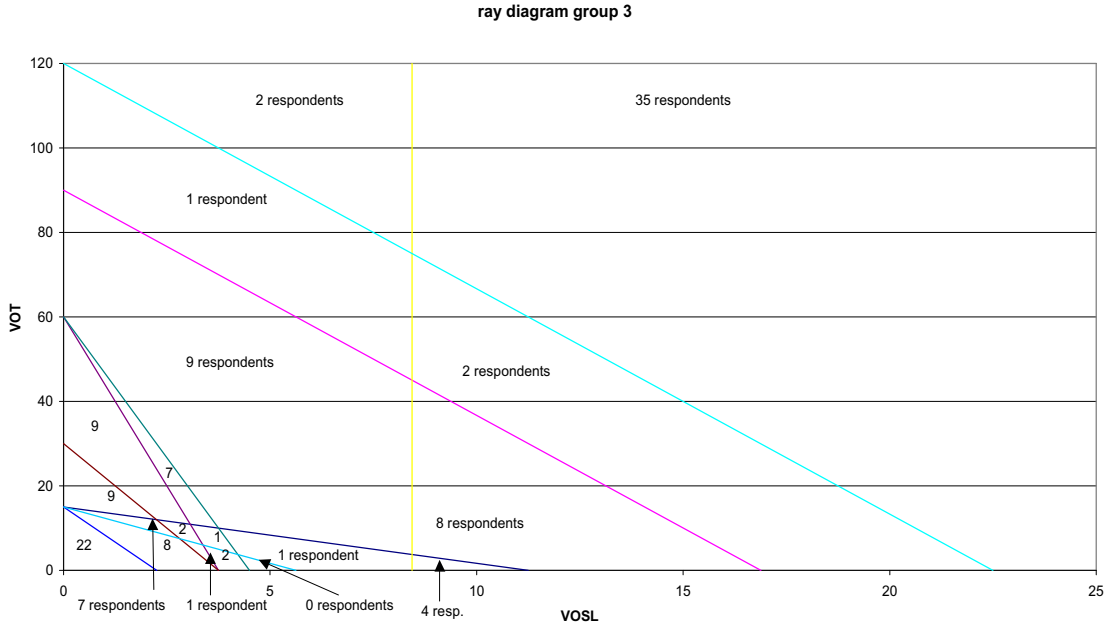


a) Group 1

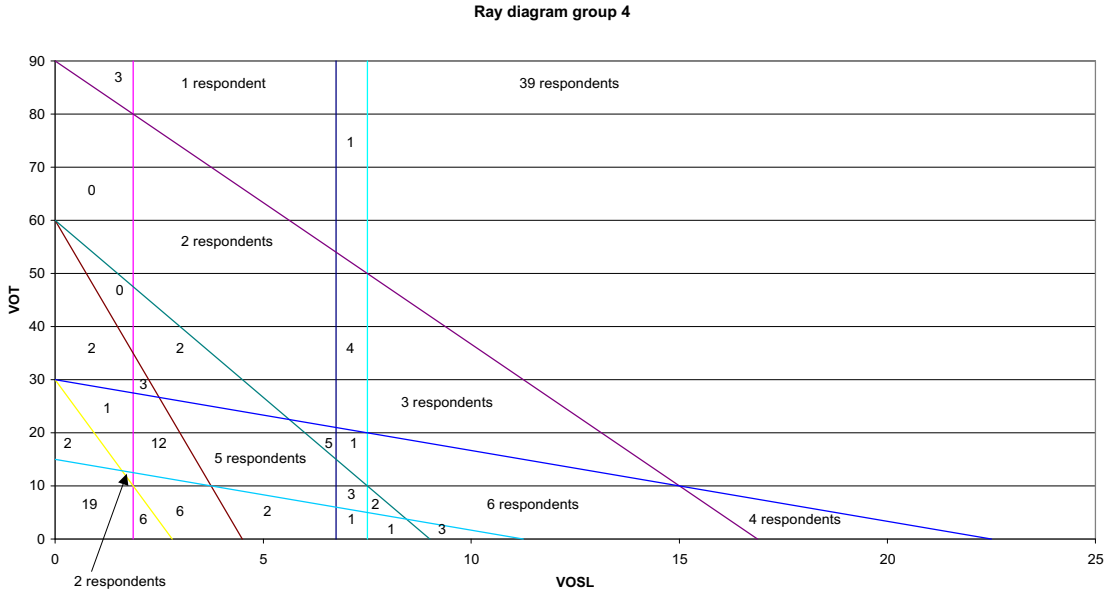


b) Group 2

Figure 1 Ray diagrams and consistent choices

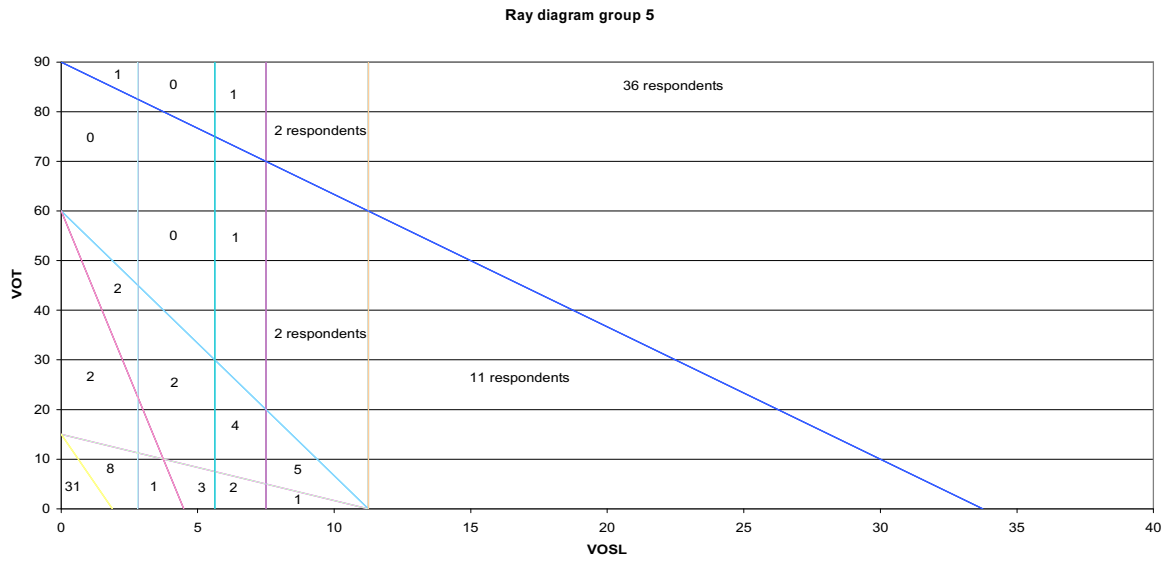


c) Group 3



d) Group 4

Figure 1 (continued) Ray diagrams and consistent choices



e) Group 5

Figure 1 (continued) Ray diagrams and consistent choices

Table 3 Inconsistent choice patterns

n	1	2	3	4	5
1	68	62	72	60	47
2	17	15	15	16	16
3	5	3	6	3	3
4 or more	0	0	0	0	0
Total	90	80	93	79	66

Note. *n* is the minimum number of choices that must be different in order to make the respondent's choice pattern consistent.

Removing all respondents with inconsistent choices would leave us with a sample in which lexicographic choice behavior is rather common. The graphs shown in Figure 1 suggest that the distribution of the *vosl* and the *vot* in the remaining observations has unexpected properties. It was already noted that the simultaneous distribution of these two variables does not appear to be unimodal. Removing the lexicographic respondents implies a truncation of the sample that may result in biased estimates of the coefficients. Moreover, a consequence of removing both the inconsistent lexicographic choices would be that less than 50% of our respondents could be used. Two conclusions may be drawn. The first is that removing the inconsistent respondents does not appear to lead to a sample that is evidently more reliable than the original one. Although many respondents have a choice pattern that is overall inconsistent, the amount of inconsistency appears to be limited and it seems highly likely that these choice patterns contain useful information. On the other hand, the subsample of consistent respondents contains a large share of respondents who have possibly answered the questionnaire in a way that prevents inconsistencies but was be caused by a superficial interest in the questionnaire, namely by giving lexicographic answers. If the lexicographic respondents are also removed, we are left with a relatively small sample with a truncated distribution of the *vot* and the *vosl*. It seems therefore better to deal with

the inconsistencies and lexicographic choices by trying to find a sensible way of analyzing the whole data set, than by removing all observations that do not pass the possibly less useful checks on consistency and lexicographic behavior. That is the approach that will be followed in the remainder of this paper.
Second, Figure 1 suggests that there is considerable heterogeneity in the *vot* and the *vosl* of the various respondents. It seems therefore important to take this into account in the analysis that follows. However, before we do this we will first in the next section look at some simpler logit models that assume that all respondents have basically the same preferences. In section 6 we reintroduce heterogeneity.

5 A basic logit model

In this section we take a different approach towards the analysis of our data. It is now assumed explicitly that the choice process contains random errors. We estimate a simple logit model and use it to derive testable predictions with respect to the occurrence of such choice patterns. These predictions are compared with observed choices.

The basic model

We start with what may be the simplest logit model that can be used to analyze the data at hand. We assume that all respondents evaluate the three attributes in exactly the same way (i.e. the coefficients in (3) are identical for all respondents) and that their utility function is linear in these attributes. Since this model is the simplest one that we will use, we refer to it as the basic model. Parameter estimates are reported in Table 4. They are all negative (as expected) and significantly different from 0. Moreover the implied *vot* and *vosl* appear to be reasonable. The inconsistencies and lexicographic responses clearly do not result in puzzling or apparently wrong estimation results. This may be so because model (5) is a reasonable representation of the actual choice process.
If we take as our working hypothesis that the estimated model is the true one, we can make some inferences into the probability of inconsistent choices and, more general, about choices that are not in accordance with the consumer's preferences as given by the estimated parameters.

Table 4 Estimation results for the basic model

Variable	Estimate	Standard error
Toll	-0.2199	*0.0064*
Fatalities	-0.0586	*0.0022*
Travel time	-0.0704	*0.0046*
Loglikelihood	-6580	

Implications for inconsistent choices

The probability that the respondent chooses alternative 1 when answering questions 2 and 10 can, on the basis of the estimated coefficients, be computed for each group. These probabilities are given in column 2 of Table 5. The predictions with respect to inconsistent and consistently right and wrong choices implied by these probabilities are given in the last three columns of this Table. They are compared with the observed frequencies of these three possibilities.
Table 5 shows – somewhat surprisingly – that the number of inconsistent choices is *smaller* than predicted by the model. The number of respondents choosing consistently the alternative that they 'really' prefer is larger than predicted for all five groups. Both consistent choices occur more

frequently than the inconsistent choices, which is also in contrast with the prediction of the model.

Table 5 Actual and predicted frequencies of responses to 2 and 10

Group	Choice Prob		Inconsistent	Two times 0	Two times 1
1	0.68	Predicted	0.44	0.10	0.45
		Actual	*0.16*	*0.20*	*0.64*
2	0.75	Predicted	0.38	0.06	0.56
		Actual	*0.16*	*0.14*	*0.70*
3	0.30	Predicted	0.42	0.48	0.09
		Actual	*0.09*	*0.65*	*0.26*
4	0.70	Predicted	0.42	0.09	0.49
		Actual	*0.19*	*0.17*	*0.67*
5	0.60	Predicted	0.48	0.16	0.36
		Actual	*0.14*	*0.34*	*0.52*

If the predicted probabilities of the logit model are correct, it is extremely unlikely that we will observe the actual frequencies of consistently right, inconsistent and consistently wrong answers. A simple test can be developed on the basis of the observation that the number of individuals with inconsistent choices is approximately a normally distributed variable with expected value n_g*p_g and standard deviation $(n_g*p_g*(1-p_g))^{.5}$, where n_g denotes the number of respondents in group g and p_g the probability that a respondent in that group g chooses the first alternative. For the observed frequencies similar computations can be carried out with similar outcomes. On the basis of these tests we have to reject the hypothesis that the logit model of Table 4 generates the observed pattern of consistent and inconsistent choices.

Implications for seemingly lexicographic choices

There are three possibilities for lexicographic choice behavior: the importance attached to each of the three characteristics may dominate that attached to the two others.[10] However, it was noted in section 2 that the scores on the three aspects are related to each other: (1) if alternative 2 has a better score on toll, its score on the number of fatalities is worse and vice versa, (2) if alternative 2 has a better score on fatalities, it always has a better than or equal score on travel time (and vice versa). Equal scores occur only for travel time. This implies that someone who made lexicographic choices with respect to the number of fatalities also made lexicographic choices with respect to travel time. The converse is not true: someone who makes lexicographic choices with respect to travel time has been asked at least once to compare two alternatives that did not differ in travel time.

The predicted frequency of lexicographic choices is the product of the probabilities of choosing the alternative with the best score on the characteristics concerned in each relevant case. For toll and fatalities all cases are relevant, for travel time the number of relevant cases depends on the group to which the respondent belongs. Table 6 compares these predictions with observed choice frequencies.

[10] It is unlikely that toll, travel time and the number of fatalities are valued positively.

Table 6 Actual and predicted frequencies of lexicographic choices

Group		Toll	Fatalities and Travel time	Travel time
1	Predicted	0.00096	0.00014	0.00048
	Actual	*0.07692*	*0.12981*	*0.25481*
2	Predicted	0.00159	0.00009	0.00107
	Actual	*0.06787*	*0.19910*	*0.20814*
3	Predicted	0.00047	0.00039	0.00058
	Actual	*0.10377*	*0.16509*	*0.17453*
4	Predicted	0.00084	0.00042	0.01124
	Actual	*0.08796*	*0.18055*	*0.19444*
5	Predicted	0.00064	0.00045	0.01308
	Actual	*0.15271*	*0.17734*	*0.20689*

The table shows that the observed frequency of lexicographic choices is much larger than predicted by the logit model. Indeed, the difference is so large that a statistical test seems superfluous. Predicted and actual frequencies of lexicographic choice behavior on travel time are always larger than or equal to those in with respect to fatalities, which is in accordance with the discussion given above. The observed frequencies are equal for group 3, which is the one in which respondents are only once confronted with a choice situation in which the travel times of the two alternative routes are equal to each other.

Two explanations

There are two possible explanations for the findings of this section. The low frequency of different choices for questions 2 and 10 may be caused by the fact that respondents remember their second choice when making the tenth.[11] In order to incorporate this possibility in the model, we simply add the second choice as an explanatory variable in the logit expression for the tenth choice.[12] We do this by adding a new variable, called *c2*, to the linear utility function (7) when it refers to the tenth choice. If alternative 1 was preferred in the second choice, *c2*=1, otherwise *c2*=-1. Estimation results of this alternative model are presented in Table 7 in the column indicated with 'memory'. The coefficient for choice 2 (=*c2*) is positive and significant as expected. There are only small changes in the other coefficients, and there is a large increase in the loglikelihood.

With respect to the lexicographic choices we may hypothesize that there are latent classes of respondent with lexicographic preferences referring to each of the three characteristics.[13] Let π_i be the probability that a respondent is lexicographic with respect to the i-th characteristic. We define *d1* as the 0-1 variable that indicates that a respondent made lexicographic choices with respect to toll and *d2* as the analogous variable that indicates lexicographic choices with respect to the number of fatalities. A third indicator, *d3*, equals 1 if the respondent made lexicographic

[11] Note that this does not necessarily imply that the respondents made a choice that is in accordance with their preferences. It is only assumed that they want to be consistent in their choice behavior. See Ariely et al. (2003).

[12] Inclusion of a lagged dependent variable does not induce inconsistency in estimation since the errors in the choice process are assumed to be independent for each choice. See Train (2002).

[13] The latent class approach is frequently used in marketing, see e.g. Kamakura and Russell (1989) and Wedel and De Sarbo (1994). The model used here in which the classes refer to respondents always choosing an alternative that is best with repsect to one characteristic is also related to discrete choice models with misclassification, see Hausman et al. (1998).

choices with respect to travel time, but not with respect to the number of fatalities. A respondent may report lexicographic choices because she belongs to the class characterized by that behavior, but also because she makes choices according to the basic logit model which by chance happen to be lexicographic. The respondents that belong to the class making lexicographic choices with respect to travel time are supposed to act in accordance with the logit model for those choices in which the two alternative do not differ in travel time. We define A as the set of the choice situations that do not differ in travel time and let p_i denote the probability of the i-th choice made by the respondent implied by the logit model. The likelihood l of the choices made by the a respondent can now be defined as:

$$l = d1\,\pi_1 + d2\,\pi_2 + (d2 + d3)\,\pi_3 \prod_{i \in A} p_i + (1 - \pi_1 - \pi_2 - \pi_3) \prod_{i=1}^{10} p_i \qquad (9)$$

If all π_is are equal to 0, this reduces to the likelihood of the basic logit model. Estimation results of this model with lexicographic classes are also reported in Table 7, in the columns indicated as 'Lex. Classes'. We find significant coefficients for all three classes. The largest lexicographic class are respondents who always choose the alternative that is safest. Approximately 10% of the respondents always choose the alternative with the lowest toll. There is also a small class of respondents that is lexicographic with respect to travel time. The coefficients for toll, number of fatalities and travel time in the logit model are now larger. These results are close to those reached when the lexicographic respondents are removed from the sample. There is a substantial increase in the loglikelihood.

The two changes in the model specification can also be combined. Estimation results for this alternative specification are reported in the last two columns of Table 7. The coefficient for choice 2 now has a smaller value. The reason is that the respondents belonging to the lexicographic classes all make identical choices in situations 2 and 10. Estimation results suggest that this effect of the lexicographic classes is insufficient to explain the low frequency of different choices in situation 2 and 10 completely.

Table 7 Estimation results for some extensions of the basic model

Model →	Memory		Lex. Classes		Mem.+Lex. Cl.	
Variable ↓	Estimate	S.e.	Estimate	S.e.	Estimate	S.e.
Toll	-0.214	*0.007*	-0.356	*0.007*	-0.347	*0.0079*
Fatalities	-0.0561	*0.002*	-0.0810	*0.002*	-0.0778	*0.0027*
Travel time	-0.0689	*0.005*	-0.0868	*0.005*	-0.0861	*0.0057*
Choice 2	2.36	*0.11*			1.93	*0.11*
Lex. Toll			0.0972	*0.009*	0.0967	*0.0091*
Lex. Fat.			0.170	*0.012*	0.170	*0.012*
Lex. Trav. time			0.0298	*0.006*	0.0274	*0.0063*
Loglikelihood	*-6229*		*-5102*		*-4931*	

The alternative model that combines the memory effect and the lexicographic classes fits the actual frequencies of lexicographic choices and different choices in situations 2 and 10 much better than does the basic logit model. This is not too surprising, given that the estimated values of π_1, π_2 and $(\pi_2+\pi_3)$ are roughly equal to the average (over the five groups) actual frequencies of lexicographic choices with respect to toll, number of fatalities and travel time, respectively. Moreover, the inclusion of the memory effects reduces the discrepancy between actual and predicted frequency of different choices in the two identical situations. This is shown in Table 8.

In computing the predicted frequencies the respondent's choice in situation 2 is used in order to compute the probability of the choice made in situation 10. It is clear from the Table that especially the frequency of choosing two times the alternative with the highest probability of being chosen in situation 2 is predicted much better now. The frequency of inconsistent choices is now underpredicted, whereas that of choosing two times the alternative with the smallest probability of being chosen in situation 2 is overpredicted.[14]

Table 8 Actual and predicted frequencies of responses to 2 and 10 in the model with memory and lexicographic classes

Group	Choice Prob		Inconsistent	Two times 0	Two times 1
1	0.69	Predicted	0.08	0.26	0.66
		Actual	*0.16*	*0.20*	*0.64*
2	0.78	Predicted	0.07	0.20	0.72
		Actual	*0.16*	*0.14*	*0.70*
3	0.18	Predicted	0.07	0.66	0.27
		Actual	*0.09*	*0.65*	*0.26*
4	0.73	Predicted	0.08	0.23	0.69
		Actual	*0.19*	*0.17*	*0.67*
5	0.59	Predicted	0.09	0.33	0.57
		Actual	*0.14*	*0.34*	*0.52*

Even though the estimation results just discussed are satisfactory in that they lead to significant coefficients and a much better fit between model and data, they are still somewhat unsatisfactory from another point of view. In the light of the results of the analysis of section 4 a major problem seems to be that the basic logit model used here assumes that all respondents have identical utility functions. This is unlikely to be true and the fact that we impose this assumption on the data probably has consequences for our conclusions with respect to answering identical questions differently and the occurrence of lexicographic choices. If the heterogeneity in our sample of respondents could be better taken into account, we might well be able to explain their choices with greater precision. This may imply that these predicted choice probabilities will often be close to 0 or 1. The predicted probability of observing different choices for identical questions will be lower for such more precisely predicted choices than it is for choice probabilities that are closer to .5. Moreover, the probability that lexicographic choices will be observed will probably also become larger, at least for some respondents. In the next section we will therefore explore the consequences of taking into account this heterogeneity for the need to introduce essentially *ad hoc* assumptions like memory-based behavior (that does not necessarily reflect preferences but only the desire to be consistent) and the existence of latent classes with lexicographic choice behavior. We concentrate especially on the frequency of different answers for questions 2 and 10. The reason has been discussed at the end of sections 2: it is unlikely that we will find an explanation for lexicographic choice behavior with a probabilistic model unless it generates choice probabilities for a non-negligible number of respondents that are very close to 0 or 1.

[14] The analogous Table for lexicographic choices is not presented because the figures are dominated by the estimated frequency of belonging to the corresponding latent class. The expected frequency of lexicographic choices for consumers acting in accordance with the logit model is of the same order of magnitude as the figures presented in Table 5.

6 The impact of heterogeneity among respondents

Incorporating observed heterogeneity among respondents

One possible way to try to improve the model is the incorporation of observed characteristics of the respondents into the model. We have done so by making the coefficients linear functions of age, gender, education and income:

$$\beta_i = \beta_{i0} + \beta_{i1}(age_i/10) + \beta_{i2}(female_i) + \beta_{i3}(lower\ ed_i) + \beta_{i4}(higher\ ed_i) + \beta_{i5}(income_i) \quad (10)$$

Education was given in seven classes and dummies were constructed for the first two (lower education) and the highest two (higher education). Income is given in classes of 1,000 Dutch guilders per year; we used class midpoints and treated this variable as if it were continuous. Estimation results are presented in Table 9. A higher age implies a smaller coefficient for toll and travel time, but a higher one for fatalities. Females attach a higher value to road safety than males. The lower and the higher educated value a toll significantly less than the reference group. The higher educated have a higher value of travel time. Respondents with a higher income attach a higher value to road safety. These results seem to be reasonable.

Table 9 Results of the logit model with respondent characteristics

	β_1 Toll	β_2 Fatalities	β_3 Travel time
Constant	-0.3841*	-0.0242*	-0.0913*
Age/10	0.0407*	-0.0047*	0.0090*
Female	0.0242	-0.0183*	0.0082
Lower ed.	0.0502*	0.0220*	-0.0006
Higher ed.	0.0331*	-0.0001	-0.0173*
Income	0.0032	-0.0028*	-0.0029
Loglikelihood		*-6437.87*	

*Coefficient is significant at p=0.05.

Table 10, which is analogous to Tables 5 and 8, shows how the predictions of the model with respect to choices in situations 2 and 10 compare with the observed frequencies. In constructing this table we had to take into account that the probability that alternative 1 will be chosen is now dependent on the characteristics of the respondent. The predicted frequencies of the various possibilities have been computed on the basis of the individual choice probabilities for all respondents belonging to each group. The indicated choice probability is therefore the average choice probability of all members of the group and the predicted frequencies of the various combinations of responses are also averages over the groups.

The predicted frequencies are always very close to the values computed for the simple logit model of the previous section. In all classes the number of predicted inconsistent choices is less than half the predicted value. Incorporating the effects of observed heterogeneity does not appear to be helpful in finding an explanation for the low frequency of different choices for situations 2 and 10. The figures for predicted and actual lexicographic choice behavior, which are not presented here, lead to a similar conclusion.

Table 10 Actual and predicted frequencies of responses to 2 and 10 in the model with observed heterogeneity

Group	Choice Prob		Inconsistent	Two times 0	Two times 1
1	0.69	Predicted	0.43	0.10	0.47
		Actual	*0.16*	*0.20*	*0.64*
2	0.78	Predicted	0.37	0.06	0.57
		Actual	*0.16*	*0.14*	*0.70*
3	0.18	Predicted	0.41	0.49	0.10
		Actual	*0.09*	*0.65*	*0.26*
4	0.73	Predicted	0.41	0.10	0.50
		Actual	*0.19*	*0.17*	*0.67*
5	0.59	Predicted	0.48	0.17	0.35
		Actual	*0.14*	*0.34*	*0.52*

The model with observed heterogeneity implies different choice probabilities for respondents with different characteristics, and this offers the possibility for a closer examination of the discrepancy between observed and actual frequencies. In Table 11 we have classified the probability that alternative 1 will be chosen in classes with breadth 0.10. The predicted probability of inconsistent choices was computed on the basis of class midpoints. The actual frequencies are also computed for each class on the basis of the predicted probability of choosing alternative 1 in situations 2 and 10. The Table shows that the actual frequencies have more or less an inverted U-shape, like the predicted frequencies, but it is not symmetric around .5.
This pattern can be visualized by carrying out a non-parametric regression of the frequency of different choices in situations 2 and 10 on the actual choice frequencies.[15] The result is Figure 2. The line for the actual frequency is based on a normal kernel. The confidence interval was determined by means of the bootstrap. The Figure shows that the difference between the actual and predicted frequencies of different answers is significant except for those individuals who have a probability of choosing alternative 1 that is close to 1.

Table 11 Predicted and actual frequencies of different choices in situation 2 and 10 for the model with observed heterogeneity

Interval	Number of observations	Actual frequency	Predicted frequency
0-.1	0	-	0.095
.1-.2	14	0.000	0.255
.2-.3	94	0.053	0.375
.3-.4	80	0.163	0.455
.4-.5	37	0.108	0.495
.5-.6	139	0.173	0.495
.6-.7	298	0.185	0.455
.7-.8	351	0.157	0.375
.8-.9	42	0.071	0.255
.9-1.0	0	-	0.095

[15] See, for instance, Blundell and Duncan (1998).

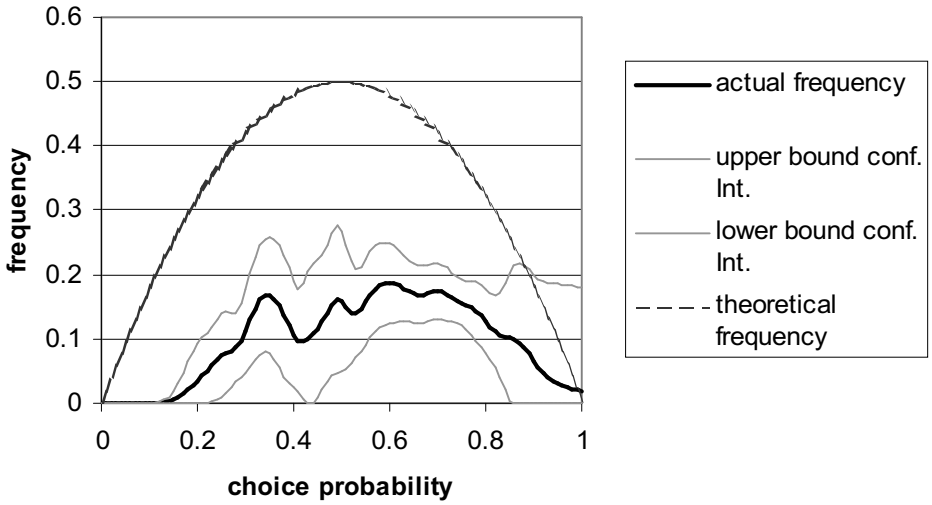


Figure 2 Non-parametric analysis of different choices in situations 2 and 10 for the logit model with observed heterogeneity

Another check on the effect of observed heterogeneity on the occurrence of different choices in situations 2 and 10 and of lexicographic choices uses the extension of the model with a memory effect and lexicographic classes. If the inclusion of respondent characteristics would help in explaining the different choices made in situations 2 and 10, one would expect other values of the parameters referring to the memory effect and the relative size of the latent classes of lexicographic respondents. Estimation results of the extended model are presented in Table 12.

Table 12 Results of the logit model with respondent characteristics, memory and lexicographic classes

	β_1 Toll	β_2 Fatalities	β_3 Travel time
Constant	-0.509*	-0.0539*	-0.139*
Age/10	0.0368*	-0.00314	0.0194*
Female	0.0181	-0.00829	0.0188
Lower ed.	0.0648*	0.0220*	-0.00584
Higher ed.	0.0204	-0.000643	-0.0167*
Income	-0.00578	-0.00282	-0.00543
Choice 2		1.90*	
Lex. toll		0.0964*	
Lex. fat.		0.169*	
Lex travel time		0.0265*	
Loglikelihood		*-4881.87*	

* Coefficient is significant at p=0.05

The values of the parameters estimated for the memory effect and the lexicographic classes are very close to those reported for the basic logit model in Table 7. This confirms the conclusion

that the inclusion of observed heterogeneity does not help in explaining lexicographic choices and the occurrence of different choices in identical situations.

Mixed logit

The fact that the number of inconsistent choices is *lower* than predicted by all the models considered thus far may be interpreted as suggesting that the errors in the true model of the decision process are not independent for consecutive choices. One can imagine that the ε in (5) can be decomposed in a part that is structural in the sense that its value differs in the population, but is not independent for consecutive choices of the same individual, whereas another part is random in both dimensions. The mixed logit model (see McFadden and Train, 2000) has such a structure.

We use the linear specification (5) as the basis for the mixed logit model to be used here. However, the parameters in this equation are now random variables. This randomness reflects differences in the value individuals attach to toll, safety and travel time. The standard logit model used earlier in this section linked such differences to observed characteristics, but now we abandon that assumption. The distribution of all the coefficients is assumed to be lognormal, with a minus sign. The lognormal distribution has been chosen in order to make the model consistent with the a priori notion that all respondents dislike all three attributes. The lognormal probability density function of the parameter β_i is:

$$g(\beta_i) = \frac{1}{\beta_i \sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(\beta_i) - \mu_i}{\sigma_i^2} \right)^2 \right) \qquad (11)$$

In order to estimate the model we should take into account explicitly that the same values of the random parameters have to be used in computing the probabilities of all choices made by the individual. Let $p^j(\beta_1,\beta_2,\beta_3)$ be the probability of the j-th choice made by an individual. The likelihood for this individual is then:

$$l_i = \int_{\beta_1}\int_{\beta_2}\int_{\beta_3}\left[\prod_{j=1}^{10} p^j(\beta_1,\beta_2,\beta_3)\right] f(\beta_1,\beta_2,\beta_3)\, d\beta_3 d\beta_2 d\beta_1 \qquad (12)$$

If all three parameters were independently lognormal distributed, f would be the product of three functions $g(\beta_i)$. However, we allow for correlation between the random coefficients, so f is a trivariate lognormal distribution.

Note that the integration in (12) is over the product of the 10 choice probabilities. This reflects our assumption that the βs are individual specific constants.[16] Part of the randomness in the respondent's evaluation of alternatives is therefore 'structural' in the sense that it reflects the tastes of the respondent. The non-structural part is reflected in the error term incorporated in the choice probabilities p^j for given values of the βs.

The model has been estimated by simulated maximum likelihood. We estimate the elements of the Choleski factor of the variance covariance matrix of the trivariate normal distribution corresponding with the lognormal density f in (12).[17] Estimation results are presented in Table 9. The diagonal elements of the Choleski factor are the standard deviations. They are all significant, indicating that there is substantial taste variation among our respondents. The last two columns of

[16] If the integral and product signs were reversed, we would also estimate a mixed logit model, but on the basis of the assumption that the random parameters in subsequent choices were independent of each other.

[17] See, for instance, Train (2002) for a discussion.

this Table present the non-diagonal elements of the Choleski factor. The only significant coefficient of these three indicates a positive correlation between the value attached to the toll and that to travel time.

Table 13 Estimation results for the mixed logit model

Variable	μ	σ		
Toll	-0.32*	1.08*		
Fatalities	-1.73*	1.46*	0.15	
Travel time	-2.09*	1.05*	0.92*	0.06
Loglikelihood	-4488			

* Coefficient is significant at p=.05.
1000 independent drawings were used for each observation.

The large increase in the likelihood suggests that this specification of this model fits the data better than those considered earlier (note that the basic logit model is the special case of the mixed logit that has all elements of the Choleski matrix equal to 0).
In order to analyze the implications of the mixed logit model for different choices in situations 2 and 10, let $p(\beta)$ denote the probability that alternative 1 will be chosen at the second or tenth question, for given values of the random parameters. The probability of inconsistent choices at the given parameters values is determined in the same way as for the basic logit model, but we should now integrate over the parameters in order to obtain the overall expected probability of such choices.
The expression we have to evaluate is:

$$2\int_{\beta_1}\int_{\beta_2}\int_{\beta_3} p(\beta,)(1-p(\beta))f(\beta)d\beta_3\, d\beta_2\, d\beta_1 \qquad (13)$$

This integral is evaluated by simulation. We use the same approach to evaluate the predicted probabilities that two times alternative 0 will be chosen and two times alternative 1. The results appear in Table 14.

Table 14 Actual and predicted frequencies of choice combinations for the mixed logit model

Group	Expected Choice Prob		Different choices	Two times 0	Two times 1
1	0.72	Predicted	0.19	0.18	0.63
		Actual	*0.16*	*0.20*	*0.64*
2	0.79	Predicted	0.16	0.13	0.70
		Actual	*0.16*	*0.14*	*0.70*
3	0.31	Predicted	0.15	0.61	0.24
		Actual	*0.09*	*0.65*	*0.26*
4	0.69	Predicted	0.17	0.22	0.61
		Actual	*0.19*	*0.17*	*0.67*
5	0.62	Predicted	0.20	0.28	0.52
		Actual	*0.14*	*0.34*	*0.52*

Comparison with Table 5 shows that the mixed logit does indeed a much better job in predicting these frequencies than the basic model. Indeed, the mixed logit model is able to explain that *both* types of identical choices occur more frequently than different choices, which stands in sharp contrast with the basic logit model.

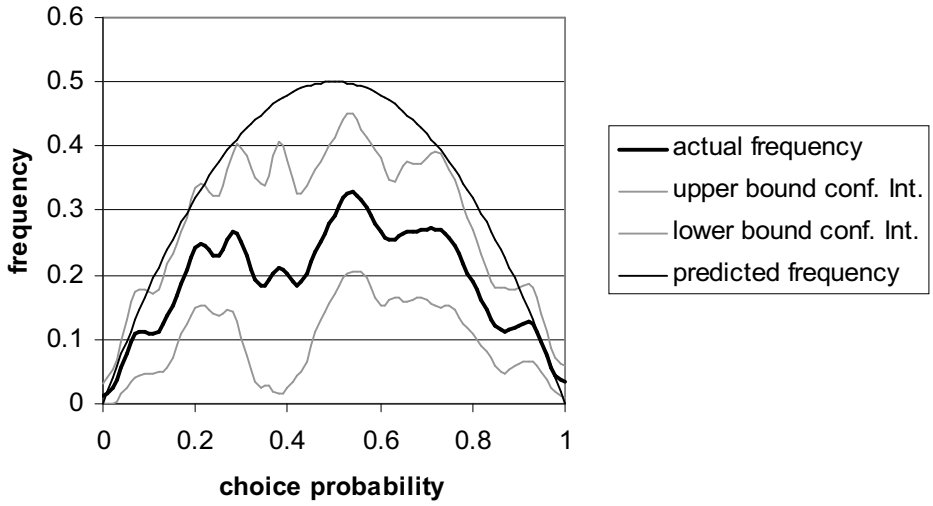


Figure 3 Non-parametric analysis of different choices in situations 2 and 10 for the mixed logit model

We have further elaborated the model's predictions about different choices by computing the posterior choice probabilities for the two alternatives in choice situations 2 and 10 on the basis of the estimated coefficients and the observed choices in situations 1 and 3-9. The method is explained in detail in Train (2002). We use these posterior choice probabilities as the basis for a non-parametric regression of the frequency of different choices in situations 2 and 10 in the same way as was done when constructing Figure 2 above. The result is Figure 3.
Comparison with Figure 2 shows that the mixed logit model does indeed a better job than the (standard) logit model with observed heterogeneity, but there still appear to be significant differences between the model predictions and the observed frequencies. It may be argued that the confidence interval shown in the Figure is too small since the posterior probabilities were treated as constants when computing this interval, whereas they are in fact expected values of random variables. However, even with a broader confidence interval it remains unsatisfactory that the non-parametric regression line is always below .35, whereas the model predicts that it should be as high as .5 when the probability of choosing either alternative is equal to .5.
The low value of the predicted frequency of different choices in situations 2 and 10 suggests that inclusion of possible memory effects will improve the performance of the mixed logit model. In order to check this, we have extended the mixed logit model with the memory effect and the latent lexicographic classes. Estimation results are presented in Table 15. We now find somewhat larger values for the μ s, somewhat smaller values for the σ s and a significant correlation between the values attached to toll and the number of fatalities. The smaller standard deviations indicate that some of the heterogeneity is now captured by the latent lexicographic classes. The parameter referring to respondents who are lexicographic with respect to travel time was put equal to zero because its value consistently became negative during the estimation procedure.

The values of the other two parameters referring to lexicographic respondents are smaller than those found in earlier models (compare Tables 7 and 12), but they are significantly different from 0 and the group that is lexicographic with respect to the number of fatalities is still of substantial size.

Table 15 Estimation results for the mixed logit model with memory and lexicographic classes

Variable	μ	σ		
Toll	-0.404*	0.897*		
Fatalities	-2.03*	0.855*	1.01*	
Travel time	-2.22*	0.143	0.843*	0.00752
Choice 2	1.73*			
Lex. Toll	0.0349*			
Lex. Fat.	0.133*			
Lex. Travel time	0**			
Loglikelihood	-4374			

* Coefficient is significant at p=.05. ** This coefficient was fixed. 1000 independent drawings were used for each observation.

7 Conclusions

In the previous sections we have analyzed SC data containing a substantial fraction of respondents that made different choices in identical situation or made lexicographic choices. Such behavior is often interpreted as indicating less reliable choices. Including these respondents in the analysis is often thought to lead to biased estimates of the *vot*, the *vosl* or other variable that are of policy interest. In this paper we take a different point of view. If it must be assumed that respondents make errors in SC-experiments, our analysis of the resulting data should take this into account. Indeed, the logit model can be interpreted as the outcome of a process in which decision-makers make mistakes. One consequence is that the logit model predicts that respondents sometimes make different choices in identical situations. The model gives precise, testable predictions about the frequency of such inconsistent behavior. Similar predictions can be derived for lexicographic choice behavior. When we compared these predictions of the logit model with the actual frequencies we found that the number of different choices in identical situations was consistently smaller than predicted, whereas the number of apparently lexicographic respondents was usually larger than predicted. It would therefore clearly be wrong to remove the respondents with different choices in identical situations from the sample since their number is actually too low.

Much of the discrepancy between model and reality can be removed by making two additions. The first is the introduction of a memory effect: respondents are hypothesized to remember their earlier choice when they are confronted with an identical situation. Since they intend to be consistent, they are more inclined to make the same choice again than would be expected on the basis of the original logit model. The second is that there are assumed to be latent classes of respondent who are lexicographic with respect to one of the aspects of the choice situations.

In the data at our disposal observable heterogeneity does not contribute to the explanation, but when the coefficients of the utility function are assumed to be random variables, part of the discrepancy between model and reality disappears. However, even in this case the memory effect

remains sizable and two of the three potentially present lexicographic classes still seem to be present, even though their estimated size now becomes smaller. One latent class disappears.
A general conclusion that emerges from the analysis is that it does not seem to be useful to remove the respondents that appear to give less reliable answers from the analysis. Such an approach may be appropriate if one uses an explanatory model that assumes that respondent make no errors at all. However, a consistency check on our data suggests that in this case one is left with a much smaller number of respondents, with many of them making (almost) lexicographic choices. Since the amount of inconsistency among the removed respondents is limited and the (nearly) lexicographic respondents are suspect for other reasons, this is an unattractive situation.
If one uses analytical that account for errors made by respondents, such as the logit model, then it makes no sense to remove all observations that incorporate errors. It would be much better to use the predictions of the model about the occurrence of such errors to check its reliability. For the data considered here this led to the conclusion that the actual number of respondents making different choices in two identical choice situations is *smaller* than predicted, whereas the actual number of lexicographic respondents is much larger than predicted. The model was then adapted so as to bring it in better agreement with the facts. Our modeling strategy thus attempts to incorporate and explain the aspects of the data that may indicate less reliable choice behavior of respondents in order to improve the overall performance of the model.

The consequences of the different model specifications for the *vot* and the *vosl* are shown in Table 16. Introduction of the memory effect and of lexicographic classes in the basic model leads to a substantial decrease in the *vot* (-22%) and the *vosl* (-16%). For the mixed logit models the expected value of *vot* and *vosl* was computed as the average over 10,000 random drawings from the simultaneous distribution of the coefficients. The *vosl* in the mixed logit model is very high if we do not explicitly take into account the lexicographic classes and the memory effect. The reason is that the presence of relatively high number of lexicographic choices, especially with respect to safety, influences the estimated distribution of the coefficients. If the extensions are made to the model, the order of magnitude of the *vosl* is the same as for the logit model without random coefficients. Indeed, the parameter σ for the random coefficient referring to fatalities is higher in the model without lexicographic classes and memory effect and the correlation with the valuation of toll is insignificant. Extending the model lead to a lower value of σ and a significant correlation with the valuation of the toll. The correlation between the valuation of toll and of travel time that is present in both varieties of the mixed logit model prevents a similar phenomenon for the value of time to occur.

Table 16 Values of time and of a statistical life in some model specifications

	Basic model	Basic model with memory and lexicographic classes	Mixed logit	Mixed logit with memory and lexicographic classes
vot	19.21	14.88	17.81	16.58
vosl ($\times 10^6$)	4.79	4.03	17.80	6.77

Units are Dutch guilders (1 euro = 2.21 Dutch guilder).

A remarkable result from our attempts to find a model that fits the data well is that respondents are more consistent in making their choices than is suggested by the standard logit model and its extension to random coefficients. There is a sizable group of respondents with lexicographic choices. Their choices may be truly lexicographic, but may also be interpreted as the result of a superficial interest in the questions posed. Moreover, there appears to be a memory effect that makes respondents more consistent in the two identical choice situations than is suggested by the logit model. Even though seven different choice situations appear between the second and the tenth, at least some respondents are more consistent in making their choices than is suggested by the logit model. A rather paradoxical outcome of our exercise is therefore that consistency, and not inconsistency, of choice behavior is the phenomenon that calls for adaptation of the conventional modeling technique.

There exist other possibilities to investigate the reliability of the data that have not been explored in this paper. For instance, the reliability of the answers given by respondents may change during the choice experiment. One can imagine that respondents learn to interpret the choice situations better as the sequence of choice situation proceeds and that their answers become more consistent with their preferences. On the other hand, one can also imagine that respondents get bored with having to answer so many questions about situations they find hard to imagine and that their answers become less reliable. In order to investigate this possibility we should allow for the possibility that the parameter σ becomes smaller or larger during the experiment. Even though we cannot estimate σ itself, we can measure changes in its value. Estimation results for this model suggest the presence of a learning effect of limited size. This effect did not contribute to the explanation of different choices in identical situations or to the occurrence of lexicographic choices.

Still another approach that may be useful for future work is to introduce an error generating mechanism that differs from that of the logit model (eq. 5). One implication of that mechanism is that the probability of making an error (a choice that does not correspond with the true preferences) is close to .5 when the utilities of the two alternative are close to each other. This implication seems to receive little support from the data. One possible reason is that respondents make a more deliberate choice when they observe that the values they attach to the alternatives are almost equal. This may explain why the regression line in Figure 3 is so much below its predicted value.

References

Ariely, D., G. Loewenstein and D. Prelec (2003) "Coherent Arbitrariness": Stable Demand Curves without Stable Preferences *Quarterly Journal of Economics* **118** 73-105.

Bates, J.J. (1994) Reflections on Stated Preferences: Theory and Practice, paper presented on the Seventh International conference on Travel Behavior, 13-16 June, Santiago, Chile.

Blundell, R. and J. Duncan (1998) Kernel Regression in Empirical Microeconomics *Journal of Human Resources* **33** 62-87.

De Palma, A., G.M. Myers and Y.Y. Papageorgiou (1994) Rational Choice under an Imperfect Ability to Choose *American Economic Review* **84** 419-440.

DeShazo, J.R. and G. Fermo (2002) Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency *Journal of*

Environmental Economics and Management **44** 123-143.
Foster, V. and S. Mourato (2002) Testing for Consistency in Contingent Ranking Experiments *Journal of Environmental Economics and Management* **44** 309-328.
Hausman, J., J. Abrevaya and F.M. Scott-Morton (1998) Misclassification of the Dependent Variable in a Discrete-Response Setting *Journal of Econometrics* **87** 239-269.
Holmes, T.P. and W.L. Adamowicz (2003) Attribute-based Methods, pp. 171-219 in: P.A. Champ, K.J. Boyle and T.C./ Brown (eds.) *A Primer on Non-Market Valuation*, Kluwer, Dordrecht.
Johnson, F.R., K.E. Mathews and M.F. Bingham (2000) Evaluating Welfare-Theoretic Consistency in Multi-Response, Stated preference Surveys, Triangle Economic Research Working Paper 0003.
Kamakura, W.A. and G.J. Russell (1989) A Probabilistic Choice Model for Market Segmentation and Elasticity Structure *Journal of Marketing Research* **26** 379-390.
Louviere, J.J., D.A. Hensher and J.D. Swait (2000) *Stated Choice Methods* Cambridge University Press, Cambridge.
McFadden, D. and K. Train (2000) Mixed Multinomial Logit Models for Discrete Response *Journal of Applied Econometrics* **15** 447-470.
Rizzi, L.I. and J. d. Ortuzar (2003) Stated Preference in the Valuation of Interurban Road Safety *Accident Analysis and Prevention* **25** 9-22.
Rosenberger, R.S., G.L. Peterson, A. Clarke and T.C. Brown (2003) Measuring Dispositions for Lexicographic Preferences of Environmental Goods: Integrating Economics, Psychology and Ethics *Ecological Economics* **44** 63-76.
Sælensminde, K. (2001) Inconsistent Choices in Stated Choice Data *Transportation* **28** 269-296.
Sælensminde, K. (2002) The Impact of Choice Inconsistencies in Stated Choice Studies *Environmental and Resource Economics* **23** 403-420.
Train, K. (2002) *Discrete Choice Methods with Simulation* Cambridge University Press, Cambridge.
Varian, H.R. (1982) The Nonparametric Approach to Demand Analysis *Econometrica* **50** 945-973.
Wedel, M. and W.S. de Sarbo (1994) A Review of Recent developments in Latent Class Regression Models, pp. 352-388 in: R.P. Bagozzi (ed) *Advanced Methods for Marketing Research* Blackwell, Oxford.