The role of preference axioms and respondent behaviour in statistical models for discrete choice

Jens Leth Hougaard
Institute of Economics
University of Copenhagen

Tue Tjur
The Statistics Group
Copenhagen Business School

Lars Peter Østerdal
Institute of Economics
University of Copenhagen

February 2004

Abstract

Discrete choice experiments are widely used in relation to health care. A stream of recent literature therefore aims at testing the validity of the underlying preference axioms of completeness and transitivity, and detecting other preference phenomena such as unstability, learning/tiredness effects, ordering effects, dominance, etc. Unfortunately there seems to be some confusion about what is actually being tested, and the link between the statistical tests performed and the relevant underlying model of respondent behaviour has not been explored in this literature. The present paper tries to clarify the notions involved and discuss what can be tested in a general frequency of choice framework and more specifically in a random utility model.

Keywords: Discrete choice experiments, Random utility model, Stated preference, Completeness, Transitivity, Statistical tests.

Address: Lars Peter Østerdal, Institute of Economics, University of Copenhagen, Studiestraede 6, 1455 Copenhagen K, Denmark. Phone: +45 35 32 35 61. Fax: +45 35 32 30 85. E-mail: Lars.P.Osterdal@econ.ku.dk.

1 Introduction

Stated preference methods are being increasingly used in many areas of applied economics, such as health care, environmental evaluation and marketing studies.

In the case of health care, for example, preferences are related to various interventions that are difficult to evaluate for the involved respondents. The choice situation is often very hypothetical and preferences usually cannot be revealed through actual behaviour and are consequently difficult to elicit. Therefore, as in any other exercise in economic modelling, the analyst is forced to make certain assumptions concerning respondents preferences in order to obtain a useful model from the available data.

The validity of such preference assumptions has been addressed by a stream of recent literature, e.g. Ryan et al. [17], Shiell et al. [21], Johnson and Mathews [6], Ryan and Bate [16], McIntosh and Ryan [12], San Miguel et al. [18], Scott [19], Ryan and San Miguel [15]. These studies aim to test whether standard properties of preferences like completeness, transitivity, stability, and (the absence of) learning and tiredness effects, ordering effects, dominated preferences, etc, are satisfied by respondents in various discrete choice experiments.

Unfortunately the literature is somewhat unclear when it comes to whether these properties are, in fact, subject to statistical tests under the given circumstances. For example, completeness has drawn much attention although it appears that within any relevant model it is meaningless to test for completeness, as we shall argue in detail.

Moreover, some of the studies refer explicitly to the random utility model, with the aim of testing the underlying preference axioms. However, the link between the statistical tests performed and the random utility model of respondent behaviour has not been explored, and it is questionable whether the proposed tests relate to a validation of this model at all.

In the present paper we try to clarify the main notions involved: We discuss how preference phenomena like completeness, transitivity, learning and tiredness effects, ordering effects, and other aspects of preference behaviour, can be given exact meaning and related to different choice models.

In Section 2 we review a few well-known results from utility theory, and recall that completeness of preferences is not a prerequisite for utility representation. We argue that completeness is merely a technical assumption that one may, or may not, impose initially but it has no real behavioural sub-

stance. Theoretically, it does play a minor role for the testable implications of utility representations, but in the present context this point is unlikely to have empirical relevance.

In Section 3 we discuss the approach of the previous literature. We focus on a recent contribution by Ryan and San Miguel [15] where tests for incompleteness and other phenomena were suggested. Their paper is in some sense representative for many of the above-mentioned studies.

In case of repeated choice, discussed in Section 4, preferences are naturally interpreted as choice frequencies (May [8]) and it becomes impossible to distinguish between "coin-flip" answers (interpreted as incompleteness) and similarity of alternatives (interpreted as indifference). Consequently, tests for completeness cannot be performed within this model. However, it is possible to test for transitivity and we suggest one way to perform such a test.

With more structure on preferences we can use the random utility model (McFadden [10]). Within the framework of this model, preferences are transitive by construction. In Section 5 we discuss the model and indicate how various preference phenomena can be interpreted and tested within extended versions of this model.

2 Completeness of preferences and utility representations

We start out by reviewing two general results from utility theory which demonstrate that completeness of preferences is not a prerequisite for utility representation and then discuss the relevance in the present context.

Let X be a finite set of alternatives, and let $\succsim$ be a binary relation on X where $x \succsim y$ has the interpretation that "x is at least a good as y". From $\succsim$ we define strict preference $\succ$ and indifference $\sim$ in the usual way.[1] It is well-known (see e.g. Fishburn [3]) that if $\succsim$ is complete (i.e. $x \succsim y$ or $y \succsim x$) and transitive (i.e. $x \succsim y$ and $y \succsim z$ implies $x \succsim z$), then there exists a function $u : X \to \mathbb{R}$ that represents $\succsim$ in the sense that

$$x \succsim y \Leftrightarrow u(x) \geq u(y). \tag{1}$$

However, it is important to recognize that completeness is not a precondition for maximization of a utility function (Peleg [14], Fishburn [3], Vind [23]).

[1] $x \succ y$ if $x \succsim y$ and $y \not\succsim x$, and $x \sim y$ if $x \succsim y$ and $y \succsim x$.

Indeed, for consistency with an underlying binary relation the utility function u only needs to ensure that dominated alternatives are not selected, i.e.:

$$x \succ y \Rightarrow u(x) > u(y). \tag{2}$$

It can be shown that there exists a utility function u satisfying (2) if and only if strict preference $\succ$ is acyclic (i.e. we never have $x_1 \succ x_2 \succ \cdots \succ x_t \succ x_1$ for finite t), see e.g. Fishburn [4].

Thus, there are two interpretations of a utility representation, (1) and (2), depending on whether completeness is assumed or not. In order to obtain a representation (1), transitivity must hold, while in (2) acyclicity must hold. Note that transitivity implies acyclicity but the converse is not true, i.e. acyclicity is the weaker property. Transitivity and acyclicity are very similar properties indeed and in practice one often seeks to reject transitivity by demonstrating cycles (see e.g. [8]). In any case, cycles, not incompleteness, seems to be the phenomenon of interest here.

Completeness is only important if we can point to intransitivities but not to any cycles in revealed preferences. For example, assume that there are three alternatives $\{x, y, z\}$. If observations indicate that $x \succ y,\ y \succ z$ but $x \not\succ z$ then if completeness is assumed we must have $z \succsim x$ and there is no utility representation (1) due to intransitivity. On the other hand, without completeness $x \not\succ z$ may indicate that x and z cannot be compared and since there are no cycles, a utility representation (2) holds.

In terms of choice experiments there is consequently no reason to assume that empirical observations are drawn from a complete ordering rather than from a partial ordering on the alternatives actually compared. For the purpose of testing representability by means of a utility function it suffices to test for cycles or intransitivity within the observed choices. As such there is no behavioural substance in the axiom of completeness; whether or not it is assumed is more or less a question of semantics.

3 What is tested in the literature?

In light of the highly limited role played by the axiom of completeness it seems surprising that recent papers (in particular in the field of health economics) are preoccupied with testing whether completeness is satisfied or not in various choice experiments, see e.g. [21] and [15] (see also [13] and [22]). Apparently, it is because they find that there is a risk that agents

when confronted with options over various alternatives (that are difficult to grasp as for example in case of health care interventions) have no well formed preferences at all and just delivers an answer to satisfy the analyst. Such behaviour is then assumed to be revealed by conflicting rankings in case of repeated choice — taken as a sign of incompleteness.

But what is in fact tested? For example, the paper by Ryan and San Miguel [15] developed a test for completeness interpreted as the assumption that individuals have "well-defined" preferences for any choice they are presented to. Unfortunately, what "well-defined" means in this context seems unclear. In the experiment, two choice situations, choice A and choice B, was repeated, with A repeated before B was introduced (this procedure was then again repeated in three waves). In each choice situation two alternatives were presented; the respondent was then asked to indicate strength of preference. If there were no clear reversals in (stated) preferences neither in the second round of choice A nor in the second round of B, this was interpreted as (an indication of) "complete preferences". If preference reversals occured both in A and B this was interpreted as "incomplete preferences". Preference reversal in A but not in B was interpreted as a "learning effect", and finally if there was a preference reversal in B but not in A then the interpretation was "random error (or "tiredness").[2]

We may try to illustrate the situation as in Figure 1 below (for simplicity a binary choice situation $\{x, y\}$ is considered with an arbitrary number of repetitions). In Figure 1A there is a preferred alternative x but random shocks may change observed choice ("random error"). In Figure 1B choices are arbitrary ("incomplete preferences"). In Figure 1C there is a learning effect in the sense that preferences converge after initial randomness ("learning").[3] Finally, in Figure 1D we have illustrated another possibility, preferences are "complete" but change over time ("unstable preferences").

Figure 1 here.

[2]Note that by chosing from the same choice sets twice, preference reversals cannot be explained by "menu-dependent" choice rules, see e.g. Sen [20].

[3]The word "learning" is somewhat misleading since there is no obvious link between learning (in the sense of becoming wiser) and convergence of choice. For example, choices may be stable due to initial ignorance and learning about the true complexity of the matter may introduce doubt and thereby unstability.

According to Figures 1A and 1B there is no point in distinguishing between incompleteness and random preferences since indecisiveness and noise cannot be disentangled based on choice observations. Hence, if the underlying model is assumed to be a random utility model (which seems sensible provided that "mistakes" are to be expected in all choice experiments) incompleteness cannot be separated from noise — this will be further clarified in Section 5. Learning effects, on the other hand, are quite different due to the fact that choices become more stable over time, i.e. noise is reduced over time. By repeating a choice once we cannot distinguish between learning and measurement error. By repeating more than once we can observe if stated preference seems to converge, see Section 5. Unstability, as in Figure 1D, seems to relate to positive autocorrelation, a rather subtile effect in this context (and difficult to identify without having data for a longer sequence of repeated choices). In particular, testing the difference between learning and unstability is impossible using tests as in [15].

The impossibility of making a clear distinction between incompleteness and noise is, in fact, very well illustrated by the empirical examples in [15]: One of the examples involves a set of questions concerning supermarket attributes, where the alternatives are only vaguely specified. For instance, prices can be "high, medium or low", without any clear quantitative specification. Obviously, many respondents will react to this by simply refusing to answer, or — as a more polite alternative — to give only vague answers. It is really a matter of taste whether this should be taken as an indication of incompleteness, an indication of noise, or an indication of alternatives that are difficult to distinguish from each other. Not surprisingly the number of imprecise preferences in the supermarket study is, in most cases, higher than in the two other studies presented, where the description of the alternatives is more precise.

The point is that completeness cannot be accepted or rejected on the basis of a standard questionnaire study, because the results of such a study will always be reported in terms of relative frequencies. If preference for x over y is defined as "a majority claims to prefer x for y", any two alternatives can be compared. The only exception is the case where all respondents refuse to answer a question. Therefore, it might be an idea as e.g. suggested in Oliver [13], to add to each question a response category labelled "comparison meaningless". If all respondents put their votes in that category we can, with some weight, conclude that either the ordering is incomplete, or the alternatives are so vaguely defined that the respondents are unable to answer.

However, this category must certainly not be confused with the mid–category labelled "indifferent", which may very well be selected as the result of a careful comparison of well–defined alternatives.

4 Testing transitivity

Assume, for simplicity, that we have a data set of pairwise comparisons (either for a given respondent facing repeated choices or a group of respondents facing a single choice). Thus, each question has the form "which of the following two alternatives x and y do you prefer?", where x and y are elements of the given set X of alternatives.[4] In addition, we make the simplifying assumption of unambiguous answers, i.e. answers like "I don't know" or "I don't care" are forbidden. As we shall see in Section 5, questions allowing for indifference can, under certain simplifying assumptions, be handled simply by exclusion of the indifference–answers from the data set.

Now, let $p(x|x,y)$ be the probability that x is chosen among alternative x and y. Since indifference is not possible, we have

$$p(x|x,y) = 1 - p(y|x,y).$$

Under our simplifying assumptions, the probabilities $p(x|x,y)$ can be estimated by the corresponding relative frequencies $\hat{p}(x|x,y)$. The preference relation induced by the choice probabilities is given by

$$x \succsim y \Leftrightarrow p(x|x,y) \geq \frac{1}{2}$$

and the estimate of this relation becomes, accordingly

$$x \hat{\succsim} y \Leftrightarrow \hat{p}(x|x,y) \geq \frac{1}{2}\,.$$

Since any pair (x,y) of alternatives will satisfy either $x \succsim y$ or $y \succsim x$, a test for completeness is meaningless. Of course, it can be argued that if $\hat{p}(x|x,y)$ is close to $1/2$ it is an indication of "coin–flip" answers, which could be explained by the respondents' lack of ability to perform a relevant comparison. But it can also be taken, simply, as an indication of x and y

[4] It is easy to generalize the method presented here to the case where three or more alternatives are presented in each comparison, see e.g. Block and Marschak [1].

being very similar alternatives, and there is no way of distinguishing between these two explanations.

However, a test for transitivity is possible. Transitivity of the induced relation $\succsim$ means (ignoring ties $p(x|x,y) = \frac{1}{2}$ which are not likely to occur) that for any triple (x,y,z) we have[5]

$$[p(x|x,y) < \frac{1}{2} \text{ and } p(y|y,z) < \frac{1}{2}] \Rightarrow p(x|x,z) < \frac{1}{2}.$$

Thus, in order to test that a *given* triple (x,y,z) does not give rise to any violation of transitivity, the following procedure will suffice:

First, check whether the three comparisons (x,y), (y,z) and (x,z) all result in *significantly decisive* conclusions. Or, equivalently, check by standard binomial tests that all three estimates $\hat{p}(x|x,y)$, $\hat{p}(y|y,z)$ and $\hat{p}(x|x,z)$ are significantly different from $\frac{1}{2}$. If this is not the case, transitivity must necessarily be accepted as far as this triple is concerned. If the three comparisons are decisive, check whether their ordering is in accordance with transitivity or not. If not (i.e. if the ordering is cyclic, which happens in 2 of the 8 possible cases), transitivity is rejected, otherwise it must be accepted.

An overall test for transitivity is, in principle, just a matter of doing this for all possible triples. But here we must (in particular if many alternatives are involved) take "mass significance" into account, i.e. the phenomenon that when many tests on level (say) 95% are performed, some of them will usually be significant just by accident. A procedure that takes this into account goes as follows:

First, isolate all pairs (x,y) for which $\hat{p}(x|x,y)$ is significantly different from $\frac{1}{2}$ on a suitable level. This level should be determined in such a way that we are almost certain that *all* these "decisive" comparisons are actually correct. Since there are $\binom{n}{2}$ pairs of alternatives (where n is the number of alternatives), the only way of ensuring this is to perform the tests on level $1 - \alpha/\binom{n}{2} = 1 - 2\alpha/(n(n-1))$, where α is chosen (as usual) to be 0.05 or 0.01 or whatever is preferred. In this way we can be sure that a "false decisive comparison" occurs with probability at most α. When these pairs have been isolated, check that the corresponding graph has no cycles. If there are cycles, transitivity is rejected, otherwise it must necessarily be accepted.

[5] Or, equivalently, *lack of transitivity* means that there exists a triple (x,y,z) (think of it as a 3–cycle $x \to y \to z \to x$) for which the three probabilities $p(x|x,y)$, $p(y|y,z)$ and $p(z|z,x)$ are all less than $\frac{1}{2}$.

It may be possible to invent more refined versions of this test, but basically there is not much more to be done. Since the hypothesis of transitivity is stated in terms of inequalities rather than equalities, it is not possible to construct a standard χ^2 test, like those usually applied in hypothesis testing for binomial data.

This test for transitivity is — though a rather elementary construction — to our knowledge not mentioned in the literature (see, however, [1] and [5] for studies of related problems). The reason for this is probably that the more structured statistical models (like the random utility model) that are usually applied in this context have transitivity as an intrinsic property. Thus, the acceptance of one of the models discussed in the next section (e.g. by an ordinary goodness–of–fit test) is an implicit acceptance of transitivity.

Notice that the relative frequency of respondents that show some sort of intransitive behaviour (as considered e.g. in McIntosh and Ryan [12]) has not much to do with this test. If the number of comparisons performed by each respondent is large, and if the alternatives are difficult to distinguish, many of the respondents are likely to get into some sort of self-contradictory behaviour. But this does not necessarily imply that the underlying relation is intransitive.

5 Statistical models for discrete comparisons

In this section it is explained how concepts like incompleteness, learning, tiredness and related issues can be discussed in the framework of a statistical model. For a detailed exposition of such models, see e.g. McFadden [11].

Discrete comparisons, in this context, refers to a situation where a number of respondents are confronted with a number of questions of the form "which of the following k alternatives do you prefer".[6]

The classical model for this kind of situations is the so-called Bradley–Terry model (see [2]), which can be stated as follows: Let π_x denote the (more or less fictive) probability that a (random) respondent, when presented to the entire set $X = \{1, \ldots, n\}$ of alternatives, answers "x". Thus, $\pi_1 + \cdots + \pi_n = 1$, provided that an answer must be given, which is assumed for the moment.

A crucial (and in some contexts questionable) assumption, called the "axiom of independence of irrelevant alternatives"[7], is that if only a subset of

[6] The case $k = 2$ corresponds to the pairwise comparison setup of the previous section.

[7] See e.g. Luce [7] and McFadden [10]

the set X of alternatives is presented to the respondent, then the probabilities can be derived from the situation involving the full set of alternatives as the conditional probabilities, given that the choice happens to fall in the subset. For example, if three alternatives x, y and z are presented, we have (with an obvious extension of the notation used in section 4)

$$p(x|x,y,z) = \frac{\pi_x}{\pi_x + \pi_y + \pi_z}\,.$$

The drawback of this assumption is that one can easily invent examples where it is unrealistic. If a pair $\{x,y\}$ of clearly distinct alternatives is extended by an alternative z which appears very similar to x, then it is not likely that the probability of selecting y will become much smaller — though this is actually what the formula suggests. Nevertheless, the model is standard in this context, and the problem described above has more to do with the interpretation of parameters than with the validity of the model in concrete situations, where the sets of alternatives involved are usually not subsets of each other, see [10].

A nice property of this model, which relates to our discussion of completeness, is that it is consistent with the simplest possible handling of "don't know" answers, in the following sense. If an indifference category – which can suitably be named 0 – is added, and if we can rely on the assumption that this alternative plays a role which is similar to any other alternative, then the "don't know" answers can be handled simply by removing them from the data set.[8]

Another nice property is automatic transitivity of the induced preference relation. Indeed, since $x \prec y$ is obviously equivalent to $\pi_x < \pi_y$, the transitivity condition reduces to the trivial statement

$$\pi_x < \pi_y \text{ and } \pi_y < \pi_z \Rightarrow \pi_x < \pi_z\,.$$

[8] For example, if two alternatives x and y are presented, the probability of choosing x when indifference is allowed becomes

$$p(x|x,y,0) = \frac{\pi_x}{\pi_x + \pi_y + \pi_0}\,.$$

But the conditional probability of selecting x, given that either x or y is selected becomes

$$p(x|x,y) = \frac{\pi_x}{\pi_x + \pi_y}$$

which according to the assumption coincides with the probability of selecting x when 0 is not among the alternatives presented.

This result is further supported by the result noticed by McFadden [10] that the Bradley–Terry model can be derived as a random utility model, i.e. a model that explains the choice made by a respondent as the one that maximizes the utility over the alternatives presented. Since choices vary from occasion to occasion and between respondents, this utility function has to be random. More specifically, let v be a function which to each alternative $x \in X$ assigns a real number $v(x)$, which can be interpreted as a sort of "average utility" in the population. The random utilities determining the choices are assumed to take the form

$$U_{ri}(x) = v(x) + \varepsilon_{xri},$$

where ε_{xri} is a random variable associated with alternative x in the i'th choice performed by respondent r. These "error terms" are assumed to be independent and identically distributed, and the choice made by a respondent in any choice situation is assumed to be the choice that maximizes the value of the random utility function U_{ri}. What McFadden [10] showed was that if the common distribution of the ε_{xri} is assumed to be the normalized extreme value distribution (c.d.f. $P(\varepsilon_{xri} \leq z) = \exp(-\exp(-z))$), then this model coincides with the Bradley-Terry model with parameters

$$\pi_{x_i} = \frac{\exp(v(x_i))}{\exp(v(x_1)) + \cdots + \exp(v(x_k))},$$

for alternatives $X = \{x_1, \ldots, x_k\}$.[9]

The random utility model has been widely applied. For example, in situations where alternatives are specified by covariate values, as in the dentist and cancer cases described in [15], a useful idea is to express the deterministic component $v(x)$ of the utility function as some specified function of a linear combination of (optionally transformed) covariate values. In this way, $v(x)$ can be split up as a sum of contributions from the covariates, which in some cases enables us to give a very concrete interpretation of covariate effects in

[9] In particular, when only pairwise comparisons are available, the model becomes

$$p(x|x,y) = \frac{\pi_x}{\pi_x + \pi_y} = \frac{\exp(v(x))}{\exp(v(x)) + \exp(v(y))} = \frac{\exp(v(x) - v(y))}{1 + \exp(v(x) - v(y))}$$

which is the standard logistic regression model for pairwise comparisons with "subtractive" logit–linear structure.

terms of prices (like the price of waiting an extra 10 minutes in a dentist's waiting room according to the average public judgement, etc.).[10]

Within the framework of the random utility model some of the more vague concepts related to discrete comparisons can be formalized. We shall try to provide some suggestions in this direction.

The fact that *incompleteness* is indistinguishable from close similarity of alternatives is clearly demonstrated by the model when a scale parameter is introduced for the error term of the random utility function. If we write the random utility as

$$U_{ri}(x) = v(x) + \sigma\varepsilon_{xri}$$

where σ is a scale parameter, similar to the standard deviation in a regression model (ε_{xri} is still assumed to be normalized extreme value distributed), it becomes clear that a large degree of incompleteness (meaning that respondents seem to give their answers more or less at random) is equivalent to a large value of σ, whereas close similarity of alternatives means that the values $v(x)$ all lie in some narrow interval. But since an upscaling of the function v is obviously equivalent to a downscaling of the error term ε_{xri}, and vice versa, it is an intrinsic property of this model that it cannot distinguish between these two phenomena. To avoid this overparametrization we may as well take $\sigma = 1$ in the model where σ is constant.

In addition, the idea of a scale parameter on the error term allows us to build a *learning effect* into the model in the following way. If respondents are exposed to the same set of alternatives several times, or to different combinations involving the same alternatives, a learning effect may be interpreted as respondents becoming more and more stable and consistent in their selections. This phenomenon becomes possible in the model if we allow for a scale parameter σ_i that varies from occasion to occasion (i). If σ_i decreases, a learning effect is present. The phenomenon that σ_i increases at some point seems to be appropriately described by the word *tiredness*.

[10] This model can be tested against the full model or the logit–linear subtractive model by standard χ^2 tests, similarly the significance of single covariates can be tested, coefficients can be estimated with standard deviations and so on. In questionaire designs where the alternatives are specified in such a way that alternatives are "balanced out", in the sense that if one alternative is preferable to another on some covariates, then other covariates will compensate for this by differring in the opposite direction, this kind of analyses are often referred to under the name "conjoint analysis". The examples given in [15] are of this kind.

Heterogeneity between respondents can also be modelled. In practice, a realistic expectation is that the random variation from respondent to respondent is more pronounced than the variation from occasion to occasion for the same respondent. Moreover, a specific respondent may very well show a stable behaviour which is different from that of another respondent. The deterministic utility function $v(x)$ represents a kind of population average, but respondents may have individual preferences that are different from this average. A model that takes this into account could be a variance–component–type model based on a random utility function of the form

$$U_{ri}(x) = v(x) + \omega\delta_{xr} + \sigma\varepsilon_{xri}$$

where $\omega\delta_{xr}$ is an error term of the same kind as $\sigma\varepsilon_{xri}$, except that it is specific to the alternative x and the respondent r, but independent of the occasion i. Computationally, this model is difficult to handle, but conceptually this is exactly what is needed to describe heterogeneity between respondents. This model is not equivalent to a Bradley–Terry model or any other simple model. Models of this kind are usually specified with normal rather than extreme value distributed error terms.

An *indifference category* can be incorporated in the model in a simple way, which in most cases is likely to be more realistic than the ignorance–of–indifference–cases method proposed earlier. Consider for simplicity the case of pairwise comparisons. Instead of assuming

$$\text{Choice} = \begin{cases} x & \text{if } U_{ri}(x) > U_{ri}(y) \\ y & \text{if } U_{ri}(x) < U_{ri}(y) \end{cases}$$

we could assume, for some parameter $\beta_0 > 0$ which can be interpreted as the "least noticeable utility difference", that

$$\text{Choice} = \begin{cases} x & \text{if } U_{ri}(x) > U_{ri}(y) + \beta_0 \\ 0 & \text{if } |U_{ri}(x) - U_{ri}(y)| \leq \beta_0 \\ y & \text{if } U_{ri}(x) < U_{ri}(y) - \beta_0. \end{cases}$$

One might even consider models where the parameter β_0 varies from respondent to respondent, in accordance with the fact that some people are more hesitant with decisive conclusions than others. A similar model — with an additional parameter $\beta_1 > \beta_0$ to determine the threshold between "preference" and "strong preference" — can be used in situations where the

responses are given on (say) a five–point scale (as e.g. in [15]). These models are closely related to the models for discrete ordinal data described in McCullagh [9].

As a final remark we mention the possibility of taking a *preference–for–first–met–alternative* parameter into the model. The order in which alternatives are presented may influence the decision taken, typically by giving a higher probability to the alternatives presented first. For this reason, it is important to balance the questionnaires in such a way that the alternatives presented are not always given in the same order. Provided that this has been done, there is a rather straightforward way of building this into the model. In the case of pairwise comparisons, it can be done by the introduction of an extra preference–for–first–met–parameter, which is simply added to the deterministic utility function's value for the first alternative before the maximization. If the utility function is written as a linear combination of covariate values, this has the simple interpretation that the property of being presented first is an extra measure of quality (represented by a dummy covariate) with potential (positive or negative) influence on the choice. For triplewise comparisons and higher, it becomes a bit more complicated.

6 Conclusion

In this paper we have examined the possibilities for embedding tests of preference axioms within probabilistic choice models. We have in particular discussed the role of completeness and transitivity, and provided some suggestions for dealing with notions like learning or tiredness, heterogeneity, indifference categories and ordering effects within the random utility model.

As demonstrated by our investigation there seems to be good reasons to start out with the random utility model which takes both completeness and transitivity as inherent properties. Although both concepts play a theoretical role and in particular transitivity can be tested within a frequency of choice model, for most available data sets it seems unlikely that transitivity can be rejected.

Finally, there seems to be many papers that exclude data from respondents for further analysis if they do not pass all tests of a certain consistency axiom. An important message of this paper is that there is no reason whatsoever to eliminate data, as long as the observed violations are within the range of what could be expected in the relevant model.

References

[1] Block H.D., Marschak, J., 1960, Random orderings and stochastic theories of responses, in *Contributions to probability and statistics*, Olkin et al. (Eds.), Stanford University Press.

[2] Bradley, R.A., Terry, M.E., 1952, Rank analysis of incomplete block designs, *Biometrika* **39**, 324-345

[3] Fishburn, P., *Utility theory for decision making*, John Wiley & Sons, 1970.

[4] Fishburn, P., Preference structures and their numerical representations, 1999, *Theoretical Computer Science* **217**, 359-383.

[5] Kendall, M.G., Babington Smith, B., 1940, On the method of paired comparisons, *Biometrika* **31**, 324-345

[6] Johnson F.R., Mathews, K.E., 2001, Sources and effect of utility-theoretic inconsistency in stated-preference surveys, *American Journal of Agricultural Economics* **84**, 1328-1333.

[7] Luce R.D., 1959, *Individual Choice Behaviour: A Theoretical Analysis*. John Wiley & Sons.

[8] May K., 1954, Intransitivity, utility, and the aggregation of preference patterns, *Econometrica* **22**, 1-13.

[9] McCullagh, P., 1980, Regression models for ordinal data, *Journal of the Royal Statistical Society B* **42**, 109–142

[10] McFadden D., 1973, Conditional logit analysis of qualitative choice behavoir. In Zarembka, P. (Ed.) *Frontiers in Econometrics*, Academic Press.

[11] McFadden D., 1986, The choice theory approach to market research, *Marketing Science* **5**, 275-297.

[12] McIntosh E, Ryan M., 2002, Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: Implications of discontinuous preferences. *Journal of Economic Psychology* **23**, 367-382.

[13] Oliver A., 2000, Complete preferences over health states: A reply to the paper by Shiell *et al., Health Economics* **9**, 727-728.

[14] Peleg, B., 1970, Utility functions for partially ordered topological spaces, *Econometrica* **38**, 1, 93-96.

[15] Ryan M, San Miguel F., 2003, Revisiting the axiom of completeness, *Health economics* **12**, 295-307.

[16] Ryan M, Bate A., 2001, Testing the assumptions of rationality, continuity and symmetry when applying discrete choice experiments in health care, *Applied Economics Letters* **8**, 59-63.

[17] Ryan M, McIntosh E, Shackley P., Methodological issues in the application of conjoint analysis in health care, *Health Economics* 1998; **7**, 373-378.

[18] San Miguel F, Ryan M and Scott A., 2002, Are preferences stable? The case of health care, *Journal of Economic Behavior and Organization* **48**, 1-14.

[19] Scott, A., 2002, Identifying and analysing dominant preferences in discrete choice experiments: An application in health care, *Journal of Economic Psychology* **23**, 383-398.

[20] Sen A., 1993, Internal consistency of choice, *Econometrica* **61**, 495-521.

[21] Shiell A, Seymour J, Hawe P, Cameron S., 2000, Are preferences over health states complete? *Health Economics* **9**, 47-55.

[22] Shiell A, Seymour J, Hawe P, Cameron S., 2000, Will our understanding of completeness ever be complete? *Health Economics* **9**, 729-731.

[23] Vind K., 2003, *Independence Additivity Uncertainty*, Springer.

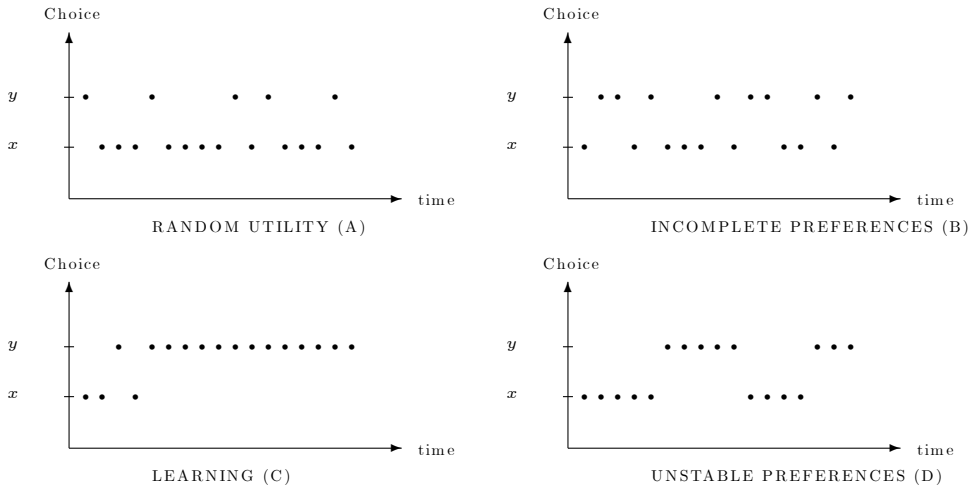
Choice
y
x
time
RANDOM UTILITY (A)
Choice
y
x
time
INCOMPLETE PREFERENCES (B)
Choice
y
x
time
LEARNING (C)
Choice
y
x
time
UNSTABLE PREFERENCES (D)

Figure 1