On the meaningfulness of testing preference axioms in stated preference discrete choice experiments

Jens Leth Hougaard*, Tue Tjur**, Lars Peter Østerdal***
*Institute of Food and Resource Economics, University of Copenhagen
**Center for Statistics, Copenhagen Business School
***Department of Economics, University of Copenhagen

March 2010

Abstract

A stream of recent studies, in particular within the context of evaluation of health care services and public goods, have developed tests of the preference axioms of completeness and transitivity and methods for detection of other preference phenomena such as unstability, learning- and tiredness effects, and random error, in stated preference discrete choice experiments. This methodological paper tries to identify the role of the preference axioms and other preference phenomena in the context of such experiments, and discusses whether or how such axioms and phenomena can be subject to meaningful (statistical) tests.

Keywords: Stated preference discrete choice experiments, Completeness, Transitivity, Random utility, Statistical tests. JEL classification: B41, C52, D01.
1 Background and introduction

Stated preference discrete choice experiments are widely used for evaluation of health states or health care services, and in other areas such as environmental evaluation and marketing- and transportation studies.

Researchers and practitioners within these areas often build statistical models of individual behavior or well-being assuming that choices are guided by utility maximization, or preferences in accordance with basic consistency conditions (known as the “preference axioms”). A stream of recent articles have indicated an increased interest in testing the preference axioms of completeness and transitivity, and identifying other preference phenomena such as unstability, learning- and tiredness, random error, etc., in the context of stated preference discrete choice experiments. A non-exhaustive list of recent studies testing such axioms in the context of health care evaluation are Ryan et al. [50], Shiell et al. [53], Ryan and Bate [49], McIntosh and Ryan [45], San Miguel et al. [51], and Ryan and San Miguel [48]. Other studies have tested these axioms in the context environmental or transport polic, such as Carlsson and Martinsson [8], Johnson and Mathews [28], Sælensminde [57][58] and others.

Tests of the preference foundations of stated preference discrete choice experiments are of considerable interest given these experiments growing popularity. Tests can also be motivated by the fact that stated preference experiments often have no real consequences for the respondents involved and hence very little can legitimately be assumed a priori about consistency of respondent behavior. Somewhat surprisingly, though, the above-mentioned studies have said little about which preference models (if any) are actually being tested and whether the preference axioms and other preference phenomena mentioned are, in fact, subject to statistical testing under the given circumstances. In particular, several studies refer, more or less explicitly, to the random utility model (with the aim of testing its underlying preference axioms), but the actual link between the tests performed and the random
utility model of respondent behavior has not been fully addressed.

This paper discusses the meaning of completeness and transitivity in the context of stated preference discrete choice experiments with particular reference to the above-mentioned recent contributions, and provides some specific suggestions for distinguishing preference phenomena like random error, learning or tiredness, heterogeneity, indifference categories, etc.. We do so by considering how the preference axioms and other preference phenomena can be specified in relation to distinct modelling frameworks, and consider whether and how these can be subject to (statistical) tests using methods which have, in fact, been applied in other strands of the literature or could be adapted to the problems at hand.

Section 2 begins by recalling some fundamentals concerning the role of the axiom of completeness for (tests of) the utility maximizing hypothesis, arguing that, within a measurement theoretical model, it is a technical assumption that may, or may not, be imposed initially, but it is unlikely to have any well-defined behavioral meaning.

In a situation with many repeated choices, preferences are naturally interpreted as choice frequencies (e.g., May [37]). When repeated choices are interpreted this way, however, 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 framework. On the other hand, tests for transitivity can be meaningful in relation to various statistical models.

Section 3 debates aspects of recent experimental studies. Tests of the completeness axiom have drawn attention although we shall argue that, within the relevant statistical models, it is meaningless to test for completeness since it cannot be disentangled from random error. The basic methodological issues we attempt to raise apply to detection of other associated preference phenomena as well. We also discuss recent experimental tests of the transitivity axiom which seem to be of a kind quite different from the above-mentioned.
In Section 4, we further connect this discussion to the well-known random utility model (e.g. Becker et al. [3], McFadden [41]). Within the framework of this much more structured model, the underlying preference relations are in a sense not only complete but also transitive by construction. But we provide an overview of how other (in some sense related) preference phenomena could be interpreted and detected within extended versions of this model. Section 5 concludes.

2 Tests of the preference axioms and the utility maximizing hypothesis

In the particular context of consumer demand under (non-)linear budget constraints, the utility maximizing hypothesis is, in theory, testable through the revealed preference axioms. There is an extensive literature on this topic and classical references include Houthakker [21], Richter [47], Afriat [1], Diewert [10], Varian [61], Matzkin [38], and Matzkin and Richter [36]. More recently, stochastic formulations and tests have been developed by Cox [9], Bandyopadhyay et al. [2], Fleissig and Whitney [16] and others.

In other lines of the literature, in particular in the literature on stated preference discrete choice experiments in the context of health and health care, tests of the utility maximizing hypothesis have focused specifically on the underlying axiom of completeness (in contrast to, for example, tests of the revealed preference axioms). The source of inspiration for conducting such tests seems to be the well-known role of the completeness axiom in measurement theory. To inform discussion, let us therefore recall the role of completeness in a basic utility model.

Suppose that the set of alternatives is denoted by $X$. To simplify matters, we will assume that $X$ is a finite set. A preference relation is a binary relation $\succsim$ on $X$.\footnote{Some authors have discussed the distinction between agents’ psychological preferences} From $\succsim$ we define strict preference $\succ$ and indifference $\sim$ in the
usual way, i.e., \( x \succ y \) if \( x \succeq y \) and not \( y \succeq x \), and \( x \sim y \) if \( x \succeq y \) and \( y \succeq x \).

It is well-known that \( \succeq \) is complete (i.e. \( x \succeq y \) or \( y \succeq x \) for all \( x, y \in X \)) and transitive (i.e. \( x \succeq y \) and \( y \succeq z \) implies \( x \succeq z \)) if and only if there exists a real-valued function \( u \) on \( X \) that represents \( \succeq \) in the sense that

\[
x \succeq y \iff u(x) \geq u(y),
\]

for all \( x, y \in X \). However, completeness is not a precondition for maximizing utility. Maximizing utility (represented by the utility function \( u \)) is consistent with an underlying preference relation \( \succeq \) whenever

\[
x \succ y \Rightarrow u(x) > u(y),
\]

for all \( x, y \in X \). There exists a utility function \( u \) satisfying (2) if and only if \( \succ \) is acyclic (i.e. there is no sequence \( x_1, x_2, \ldots, x_t \) where \( x_1 \succ x_2 \succ \cdots \succ x_t \succ x_1 \)), see, e.g., Fishburn [15].

While associating the axiom of completeness with an agent’s ability or willingness to pick an alternative in a non-forced decision situation might have some merit (see Section 3 below for further discussion), it has little to do with tests of the utility maximizing hypothesis, for two reasons. First, even if completeness is violated, choices made may well be rationalizable by a utility function in the sense of (2). Second, the option “unable to decide” is not a real alternative, i.e. it is not an element of \( X \) (that would, at least, constitute a rather odd modelling framework). It is therefore questionable whether the inclusion of such an option has any meaning in the particular context of tests of either of the utility maximizing models given a set \( X \) of alternatives of interest. For these or related reasons, the use of forced choices and their choice behavior or “revealed preferences”, see, e.g., Mandler [39]. In the context of stated preference discrete choice experiments, \( \succeq \) would probably refer to psychological preferences between pairs of alternatives, i.e. \( x \succeq y \) means that the respondent believes he or she is at least as well off with \( x \) as with \( y \). In a choice context, \( x \succeq y \) could mean that there is a choice set that contains \( x \) and \( y \) in which \( x \) is (sometimes) chosen.
is indeed common in experiments (as it is also the case in the experiments discussed in Section 3 below).

Assuming completeness, utility maximization in the sense of (1) holds if transitivitiy holds, while in (2) acyclicity must hold. Transitivity implies acyclicity but the converse is not true. Intransitivity and cycles are however closely related preference phenomena and in practice one often seeks to find violations of transitivity by means of cycles; see, e.g., May [37]. Thus, cycles (or intransitivity for that matter) and not the question of completeness is the behavioral phenomenon of interest here (see also Mandler [39]). In relation to tests of the utility maximizing hypothesis, it therefore suffices to test for cycles or intransitivity. Whether or not the axiom of completeness is assumed is more or less a matter of taste.\(^2\)

Since inconsistency in binary choices essentially is related to intransitivities (or cycles) we shall briefly consider the basic ideas behind such tests on a model of repeated choices. Assume, for simplicity, that we have a data set of pairwise comparisons for a given respondent facing repeated choices between alternatives in a given set \(X\).\(^3\) Options like “I don’t know” or “I am indifferent” are not accepted.

Now, let \(p(x|x, y)\) be the probability that \(x\) is chosen among alternatives \(x\) and \(y\). The probabilities \(p(x|x, y)\) can be estimated by the corresponding

---

\(^2\)Eliaz and Ok (2006) recently proposed a choice model, where the agent may choose more than one alternative, that allows us to distinguish between an agent being indifferent vs. indecisive on the basis of her observed choice behavior. According to this model, if both \(x\) and \(y\) are chosen in the binary choice situation involving \(x\) and \(y\), the agent is “indifferent” between these two alternatives if she treats \(x\) and \(y\) identically in all choice situations that involve these two alternatives — otherwise the agent is “indecisive” (for details, see Eliaz and Ok, 2006). Their model is deterministic, but the feature that choices are set-valued might be interpreted as the agent randomizes (with unspecified subjective probabilities) between the alternatives. The model is noteworthy as it offers a distinction between indifference and indecisiveness which is, in principle, testable. However, it remains unresolved how to distinguish empirically when choices are single-valued and whether there is a need for distinguishing between choices and psychological preferences.

\(^3\)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., Marschak [35] and Block and Marschak [6].
relative frequencies \( \hat{p}(x|x, y) \). The preference relation induced by the choice probabilities is given by

\[
x \succ y \iff p(x|x, y) \geq \frac{1}{2},
\]

and accordingly the estimate of this relation, \( \hat{\succ} \), becomes

\[
x \hat{\succ} y \iff \hat{p}(x|x, y) \geq \frac{1}{2}.
\]

Since any pair \((x, y)\) of alternatives will satisfy either \( x \succ y \) or \( y \succ x \), 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 \( y \) being very similar alternatives, and there is no way of distinguishing between these two explanations.

A test for transitivity is possible. The induced relation \( \succ \) satisfies *weak stochastic transitivity* (Fishburn [14] surveys theories of stochastic transitivity) if for any triple \((x, y, z)\) we have

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

In order to test that a *given* triple \((x, y, z)\) does not give rise to any violation of transitivity, we could check whether the three comparisons \((x, y)\), \((y, z)\) and \((x, z)\) all result in *significantly decisive* conclusions. That is, by standard binomial tests we could check whether 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 the three comparisons are decisive, and their orderings are not in accordance with transitivity (i.e. if the ordering is cyclic, which happens in 2 of the 8 possible cases), transitivity for the triple \((x, y, z)\) is rejected, otherwise it is not.

However, an overall test for transitivity is not just a matter of doing this
for all possible triples. One reason is that (in particular if many alternatives are involved) we need to 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. While it is possible to correct for this (applying a Bonferroni correction of the significance level), such tests have attracted little attention in the literature.\textsuperscript{4} The reason for this is probably that they have little statistical power and the more structured statistical models (like the random utility model) that are usually applied in this context have transitivity as an intrinsic property.\textsuperscript{5} Thus, the acceptance of one of the models discussed in Section 4 below (e.g. by an ordinary goodness-of-fit test) is an implicit acceptance of transitivity. More powerful tests of overall transitivity have, however, been invented. A test for transitivity was employed by Tversky \cite{60} in an experimental study, where the alternative to transitivity was more narrowly specified by the nature of the experiment.\textsuperscript{6} Iverson and Falmagne \cite{22} have a sophisticated study of the distributions of the statistics related to this kind of hypothesis testing. See also Tsai and Böckeholt \cite{59} for a more recent study.

\textsuperscript{4}See, however, Kendall and Babington Smith \cite{23} and Block and Marschak \cite{6} for studies of related problems.

\textsuperscript{5}A data set of pairwise comparisons could also be collected from a group of respondents and we can imagine that the experimenter wants to test if it is possible to interpret the choices as if they were all generated by a single “representative” individual with stochastic transitive preferences. In this case it is more likely that a sufficiently large data set is available for statistical testing. Problems of this kind might have interest from, e.g., a social choice perspective. It is well-known, however, that transitivity of individual preferences do not necessarily imply a transitive preference relation for the representative individual and vice versa (e.g. May \cite{37}).

\textsuperscript{6}In a more structured framework involving choice over risky prospects, studies of the specific patterns of violations of transitivity have been used to find support for regret theory (e.g., Loomes et al. \cite{29,30} and for prospect theory (e.g., Starmer \cite{55}).
3 What is tested in recent studies?

In light of the very limited role of the axiom of completeness it may seem surprising that recent papers are preoccupied with testing whether completeness is satisfied or not in choice experiments, see, e.g., [53] and [48] (see also [46] and [54]). Apparently, it is because experimenters find that there is a risk that respondents when confronted with various alternative options (that are difficult to grasp as for example in case of health care interventions or environmental evaluation) have no well-formed preferences, but still try to deliver an answer in order not to disappoint the experimenter (forced answers). Such behavior is then assumed to be revealed by conflicting rankings in case of repeated choice — taken as a sign of incompleteness.

But does it make sense to test for completeness in a model where choices are forced? Ryan and San Miguel [48], for example, develop a test for completeness interpreted as the assumption that individuals have what they call “well-defined preferences” for any choice they are presented. In the experiment, two specific choice situations, choice A and choice B, were both repeated during the experiments, with choice A repeated before choice B was introduced (and then later repeated).

In each choice situation, two alternatives a and b were presented and the respondent was asked to select one of the following options: 1) “Strongly prefer a” 2) “Prefer a” 3) “Indifferent” 4) “Prefer b” 5) “Strongly prefer b”. If no reversals in stated preferences neither in the second round of choice A nor in the second round of B was observed, Ryan and San Miguel interpret this as (an indication of) “complete preferences”. If preference reversals occurred 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

\[ \text{The whole procedure was then again repeated in three waves.} \]

\[ \text{Certain changes in stated preferences, such as a change from “strongly prefer a” to “prefer a”, was not counted for a preference reversal.} \]
“tiredness”).

We can get an intuitive grasp of the preference phenomena that Ryan and San Miguel and others may have had in mind using Figure 1, where we have also added a fourth possibility, “unstable preferences” (that shall be explained shortly). Figure 1 illustrates, for simplicity, a binary choice situation with $X = \{x, y\}$ where the choice situation is repeated many times. Figure 1A illustrates a situation where there is a preferred alternative $x$ but random shocks occasionally change observed choice (and can be called “random error”). Figure 1B illustrates a situation, where $x$ and $y$ cannot be meaningfully compared and the choices — forced through by the analyst — are arbitrary (“incomplete preferences”). In Figure 1C there is a learning effect in the sense that preferences seemingly converge after initial randomness (“learning”). 9 Finally, in Figure 1D we have illustrated another possibility, preferences are well-articulated at each point in time but change over time (“unstable preferences”) — possibly triggered by some observed events.

9The word “learning” may be imprecise since learning (in the sense of becoming more well-informed or wiser) is not necessarily the same as convergence of choice. Choices may, for example, initially be stable due to ignorance and gradual learning about the true complexity of things (or preferences for diversity) may introduce doubt — and thereby unstability.
As Figures 1A and 1B illustrates, there is no point in distinguishing between incompleteness and random error since indecisiveness and noise cannot be disentangled based on such choice observations.\textsuperscript{10} Hence, if the underlying model is assumed to be a random preference/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 discussed in Section 4. 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 it many times we can observe if stated preference seems to converge, see Section 4. Unstability, as in Figure 1D, could be identified with positive autocorrelation between successive choices, but also here it takes many repetitions to disentangle this effect from other effects. In particular, testing the difference between learning and unstability is impossible using tests as in [48].\textsuperscript{11}

\textsuperscript{10}Note that by choosing from the same choice sets twice, preference reversals cannot be explained by “menu-dependent” choice rules, see, e.g., Sen [52].

\textsuperscript{11}The impossibility of making a clear distinction between incompleteness and noise is, in fact, very well illustrated by the empirical examples in [48]. One of the examples involves a set of questions concerning supermarket attributes, where the alternatives are
Completeness interpreted as having well-formed preferences 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 “in a majority of cases \( x \) is preferred to \( y \)”, any two alternatives can be compared. The only exception is the case where respondents can refuse to answer a question. Therefore, it might be an idea as, e.g., suggested in Oliver [46], 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.

Concerning tests of transitivity, some studies (e.g., Sælensminde [57][58]) state that the transitivity axiom is tested using so-called “ray diagrams” but it is not explicitly specified what is meant by transitivity in this context, and it is unclear whether the consistency tests performed relate to a test of this axiom or other preference phenomena (such as straight-line indifference curves).

Other studies, e.g., Carlsson and Martinsson [8] and McIntosh and Ryan [45] use a random utility model but seem to disregard the stochastics when they test for transitivity with the purpose of providing internal validation of their framework. These studies are concerned with the absolute or relative 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.
number of respondents that show some sort of intransitive behavior on at least one occasion. 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 behavior. But this does not necessarily imply that there is an underlying preference relation (in a relevant model of respondent behavior) which is intransitive.

Recent experimenters have also been concerned with the idea of exclusion of data from respondents who do not pass all “consistency checks”. The choice models suggest 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 statistical model: While experimenters should be free to interpret single deviations from “consistency” as they want, it would be wrong to associate such consistency checks with validations of the relevant discrete choice models or the use of such models for utility assessments in welfare studies.\textsuperscript{12}

4 Testing preference axioms (or other preference phenomena) in the random utility framework

In this section we discuss how concepts like incompleteness, learning, tiredness, and related issues can be analyzed in the framework of more structured statistical models. For a general treatment of such models, see, e.g., McFadden [42].

Discrete comparisons, in this context, refers to a situation where a respondent is confronted with a number of questions of the form “which of the following \( k \) alternatives do you prefer”.\textsuperscript{13} The classical model for this kind of

\textsuperscript{12}Lancsar and Louviere [25] have recently, and independently of an earlier version of this paper [20], made a somewhat similar point.

\textsuperscript{13}The case \( k = 2 \) corresponds to the pairwise comparison setup of the previous section.
situations is commonly called the Bradley-Terry-Luce (BTL) model, which can be stated as follows: Let $\pi_x$ denote the (more or less fictive) probability that a 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.

A crucial (and in some contexts questionable) assumption, called the “axiom of independence of irrelevant alternatives” (e.g., Luce [33], McFadden [41]), is that if only a subset of 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 2)

$$p(x|x, y, z) = \frac{\pi_x}{\pi_x + \pi_y + \pi_z}.$$ \[14\]

A noteworthy property of the BTL model is that it is consistent with the simplest possible handling of “don’t know” or obstructive answers, in the following sense. If an “undetermined” category — which can suitably be named 0 — is added, and if we can rely on the modelling 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.\[15\] As indicated previously, it is, however, questionable whether such a modelling approach is likely to make much sense.

\[14\] 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 has been widely applied.

\[15\] For example, if two alternatives $x$ and $y$ are presented, the probability of choosing $x$ when indifference is allowed becomes

$$p(x|x, y, 0) = \frac{\pi_x}{\pi_x + \pi_y + \pi_0}.$$
Another noteworthy property of the BTL model is automatic transitivity of the induced preference relation. Indeed, since $x \prec y$ is equivalent to $\pi_x < \pi_y$, the transitivity condition reduces to the trivial statement

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

Block and Marschak [6], Luce and Suppes, [34] and McFadden [41] showed that the BTL model can be derived as a random utility model where the stochastic element has a specific parametric form. More specifically, let $v$ be a function which to each alternative $x \in X$ assigns a real number $v(x)$. Random utilities determining the choices are assumed to take the form

$$u_i(x) = v(x) + \varepsilon_{xi},$$

where $\varepsilon_{xi}$ is a random variable associated with alternative $x$ in the $i$'th choice. 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 $u_i$. Becker et al. [3] show that a large class of the Fechner models of random errors which have attracted considerable attention in the recent years (e.g. Hey and Orme [19], Loomes and Sugden [27], Loomes [26], Blavatskyy [5]) are also binary random utility models. If the distribution of the $\varepsilon_{xi}$ is assumed to be the normalized extreme value distribution (c.d.f. 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.

The random preference model offers a quite distinct stochastic specification, see, e.g., Loomes and Sugden [31]. Another distinct stochastic specification is the pure tremble model (Harless and Camerer [17]). For comparisons of different stochastic specifications (in the context of choice over risky prospects) see, e.g., Loomes and Sugden [27], Loomes [26], and Blavatskyy [5].
\[ P(\varepsilon_{xi} \leq z) = \exp(-\exp(-z)), \]
then this model coincides with the BTL model with parameters

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

for alternatives \( \{x_1, \ldots, x_k\} \subseteq X. \)

Falmagne [12], McFadden and Richter [44] and others have established necessary and sufficient conditions for theoretical choice probabilities to be consistent with random utility maximization. McFadden [43] contains many references. Stochastic tests, however, have been much less developed; though see Koning and Ridder [24] for a related study.

There is an extensive literature on tests of various types of respondent behavior in the BTL random utility framework, and we shall refrain from an attempt to survey this broad field (see, however, Hensher et al. [18], Louviere [32], and Swait and Adamowicz [56] for other recent discussions). Rather, we shall try to synthesize how some of the previously mentioned, and somewhat vaguely defined, preference phenomena related to discrete comparisons can be formalized and tested within the framework of the random utility model.

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

\[ u_i(x) = v(x) + \sigma \varepsilon_{xi} \]

where \( \sigma \) is a scale parameter, similar to the standard deviation in a regression model (\( \varepsilon_{xi} \) 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 \( \sigma \), whereas close similarity of alternatives means that the values \( v(x) \) all lie in some narrow interval. But it is well-known that an upscaling of the function \( v \) is equivalent to a downscaling of the error term \( \varepsilon_{xi} \), and vice
versa, and 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 $\sigma$ is constant (see, e.g., Ben-Akiva and Lerman [4]).

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 $\sigma_i$ that varies from occasion to occasion ($i$). If $\sigma_i$ decreases, a learning effect is present. The phenomenon that $\sigma_i$ increases at some point seems to be appropriately described by the word tiredness.

An indecisiveness category (which could be interpreted as associated with “indifference” or “incompleteness”) can be incorporated in the model in a simple way, which in most cases is likely to be more realistic than the ignorance-of-indecisiveness-cases method proposed earlier. Consider for simplicity the case of pairwise comparisons. Instead of assuming

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

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

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

This sort of choice model was estimated in Hey and Orme [19]. One might even consider models where the parameter $\beta_0$ 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 [48]). These models are closely related to the models for discrete ordinal data described in McCullagh [40].

5 Conclusion

We have distinguished three modelling frameworks: a general measurement theoretical model, a frequency-of-choice model, and a random utility models, to shed light on the challenges of identifying an appropriate modelling framework when addressing the relevance and the actual merit of (tests of) preference axioms and phenomena. It is questionable whether the tests that have been employed in recent stated preference discrete choice experiments actually relate to a validation of any of those models. In particular, we have argued that, although recent conceptual developments (e.g, Mandler [39] and Eliaz and Ok [11]) may turn out helpful in devising new ways to think about incompleteness, the empirical relevance of existing tests of this axiom remains highly limited. Transitivity can be tested for example within a frequency of choice model, although for most realistic data sets it seems unlikely that transitivity can be rejected, at least at the individual level. In this respect it seems reasonable to work with statistical models that treat these properties as inherent. However, even under these assumptions, the amount of data required for properly analyzing choice behavior at individual level is often unrealistic.

Having said that, it should be emphasized that we do not in any respect want to downplay the importance of internal and external validations of discrete choice models that are being employed. Quite to the contrary, since statistical testing is sometimes meaningless in relation to specific preference conditions and models, there is all the more reason to critically consider the
respondents’ ability to deliver useful and reliable answers in stated preference experiments.

Acknowledgements: We would like to thank, without implicating, Martin Browning, John Cairns, Søren Johansen, and Helgi Tómasson for helpful conversations and comments.

References


