Università degli Studi di Milano

Graduate School in Social, Economic and Political Sciences

Ph.D. in Economics, 22nd Cycle

Reliability of Stated Preference Methods

Thesis supervisor: Prof. Massimo Florio

Ph.D. candidate: Giovanni Perucca

December 2010

Contents.

Introduction. 5

Reliability of stated preference methods. 9
1. Introduction. 10
2. The evaluation of WTP and WTA. 11
2.1. SP reliability and ordinal utility theory. 12
2.2. SP reliability and the behavioural economics programme. 19
2.3. SP reliability: discussion. 24
3. Surveys about attitudes and perceptions. 31
4. Conclusions. 34
Notes 37
References. 38

Logical consistency in choice experiments: an application to the value of travel time. 45
1. Introduction. 46
2. Time allocation theory. 47
3. Methodology. 49
4. Survey design. 52
5. The problem of LC. 52
6. Results. 57
7.Conclusions. 60
Notes 61
Appendix. 62

Consumers' satisfaction and quality: evidence from a cross country comparison about railway transport 75
1. Introduction. 76
2. The data. 77
3. The empirical model. 79
4. Results. 81
5. Conclusions. 85
Notes 87
Appendix 1. 88
Appendix 2. 97
The data. 97
The empirical model. 98
Results. 98
Conclusions. 99
References. 105

Concluding remarks. 109

Introduction.

Many times governments and policy makers have to choose among different projects or policies to implement. In principle, the best choice is the one which maximizes the social welfare that, in turn, depends on individual preferences. But very often preferences are unknown and even not observable.

In practice, a common procedure is to directly ask a sample of individuals about their preferences, which are therefore *stated* by agents rather than *revealed* by their behaviour. Methods for preference revelation can be classified into two broad families.

The first one involves the case in which respondents are asked to simulate their market behaviour in a fictitious context designed by the researcher. The final goal of these studies is the estimation of willingness to pay (WTP), or willingness to accept (WTA), for changes in provision of non-market goods. A large literature investigates both theoretical issues connected with these procedures (Bates, 1988) and empirical results from country experiences (Mackie at al., 2003).

The second family of surveys are commonly employed in public opinion analysis. In this case respondents are asked to reveal their current attitudes, whilst in some circumstances they are required to state their satisfaction with a certain policy or service. In the last decades the interest towards such analysis largely increased and a broad amount of surveys have been systematically collected (Rabin, 2002).

Whatever the kind of analysis, when individuals correctly report the behaviour they would keep in a real context, or honestly admit their attitudes and perceptions, the target of the policy maker is reached. Hence, the issue of reliability of stated preferences becomes crucial in order to understand what we can learn from surveys and how SP analysis can be exploited by policy makers.

Our research question is simply the following one: can we trust in SP methods?
In order to answer this question the work is organised in three sections.

The first one is devoted to the definition of the concept of "reliability". In the first place, the latter depends on the family of SP methods we are dealing with.

When individuals are required to replicate their market behaviour in a fictitious scenario, two perspectives can be applied: the first one based on mainstream economic theory (Hicks and Allen, 1934) and the other one in accordance to the so called behavioural programme (Sunstein and Thaler, 2008). Both approaches are discussed, pointing out the problematic issues which characterise each methodology and trying to propose a definition for the concept of reliability.

The second family of surveys can be classified into two sub-groups, based on the object of the analysis. The first group includes all situations where agents are required to reveal their actual behaviour (Bertrand and Mullainathan, 2001) while the second one is composed by those studies in which agents are asked to express their feelings or perceptions about a certain aspect of their life (McFadden *et al.* 2005). Again, the concept of reliability has been investigated for each group of surveys.

The second and the third sections are devoted to empirical works which try, recalling the definition of reliability suggested in the first chapter, to apply this concept to empirical studies.

In the second chapter the results of a survey about WTP for travel time are presented. This SP study has been specifically carried out for this work and includes 407 interviews conducted on the regional train between Turin and Milan in June and July 2008. The design of the questionnaire is shaped based on other surveys focused on a similar environment (TRT, 2001).

The issue of reliability in WTP estimations has been discussed by a few studies (Salelensminde, 2002), which were based on a different definition of reliability with respect to the one proposed in this work. Results show how a significant share of respondents (about 25 per cent) finds difficult to properly report their market behaviour and suggest that these troubles are connected with agents' age and with the reason of the trip (work or leisure). The econometric analysis is conducted through a multinomial logit model in the first part, devoted to the evaluation of the value of ravel time. In the second part, the analysis of the relation between "unreliable" statements and individual characteristics has been performed using both an ordered logistic regression and a Bayesian network, comparing the results of the two methods.

The third section makes use of the results of three Eurobarometer surveys, carried out in 2000, 2002 and 2004 in fifteen European countries in order to investigate consumers' satisfaction with price and quality of the services. Again, according to the definition provided in the first chapter, the issue of reliability has been analysed following previous literature on this subject (Fiorio and Florio, 2008).

The econometric model applied to this section is a generalised ordinal logistic regression (Williams, 2007). Results are consistent with previous literature findings and suggest how surveys should be cautiously interpreted in order to obtain meaningful information.
Finally, the last part of the third section presents a replication of the exercise applied to the survey presented in the second chapter. Even if the first goal of that survey consisted in the analysis of travellers' preferences about time, we asked the respondents to state, on a scale from 1 to 10 their satisfaction with railway transport in terms of punctuality. The results of this last experiment seem to suggest how satisfaction depends, rather than on individual characteristics such as age or education, on travellers' attitudes and habits.

Reliability of stated preference methods

Abstract. Sometimes governments and policy makers have to choose among different projects or policies to implement. They know that different states of the world require different policy choices to be made but, at the time the policy has to be chosen, they are uncertain about the future economic environment. In order to obtain additional information they can mainly rely on two methods: revealed preferences (RP) and stated preferences (SP). The first method consists in collecting information about actual choices made by consumers: this technique tells us something about the current economic environment but almost nothing about how preferences would change in a future and still hypothetical context. In order to collect more information, a sample of consumers can be asked, to state its preferences comparing several scenarios, each of them describing an alternative and hypothetical state of the world. In this case another issue arises: how much reliable are these results? Do SP correctly replicate real preferences? This point can be analysed from two different viewpoints. On the one hand, according to traditional assumptions in economic theory about preferences and consumers' behaviour, the issue can be viewed as an information transmission problem, where respondents may have an incentive to reveal false information about the future state of the world in order to distort policy in a direction that they find beneficial. On the other hand, by relaxing certain assumptions about rationality, some respondents could not be able to answer the questionnaire in accordance to their own preferences, due for example to the difficulty of the task.

Contents.

1. Introduction. 10
2. The evaluation of WTP and WTA. 11
    2.1. SP reliability and ordinal utility theory. 12
    2.2. SP reliability and the behavioural economics programme. 19
    2.3. SP reliability: discussion. 24
3. Surveys about attitudes and perceptions. 31
4. Conclusions. 34
Notes 37
References. 38

1. Introduction.

Many times governments and policy makers have to choose among different projects or policies to implement. The best choice is the one which maximizes the social welfare that, in turn, depends on individual preferences. But very often preferences are unknown and even not observable.

A common procedure is to directly ask a sample of individuals about their preferences, which are therefore *stated* by agents rather than *revealed* by their behaviour. When individuals exactly report, in the SP exercise, the behaviour they would keep in a real context, the target of the policy maker is reached. Unfortunately, no evidence is available to discern whether or not SP and RP coincide. Then, the fundamental issue becomes the following: can we trust in SP surveys?

In order to answer this question as a first step we have to define the concept of *reliability*. The latter depends on the structure and purpose of our SP study. Two big families of surveys can be identified. The first one involves the case in which respondents are asked to simulate their market behaviour in a fictitious context designed by the researcher, whilst in the second category individuals have to report their attitudes or feelings about a certain aspect of their life.

The first group of surveys are widely used in order to estimate willingness to pay (WTP), or willingness to accept (WTA), for changes in provision of non-market goods. In this context the reliability of our results depends on the assumptions underlying consumers' behaviour. On the one hand, according to *ordinal utility theory* (Hicks and Allen, 1934) we can assume individuals to be perfectly able to choose the option they prefer. More recently, however, the so called *behavioural programme* pointed out that very often consumers fail to maximise their utility due to bounded rationality, asymmetric information, etc. As a consequence, based on our hypothesis about consumers' behaviour we can, *ex-ante,* guess how respondents are supposed to properly state their preferences and, *ex post*, check whether their statements are consistent with our assumptions.

Surveys of the second kind are commonly employed in public opinion analysis. In this case, a further distinction concerns the object of the study. Sometimes respondents are asked to reveal their current attitudes, for example how many hours they spend watching TV in a day, whilst in some circumstances they are required to state their satisfaction with a certain policy or service. Again, the concept of reliability can be defined based on this classification.
The discussion is organised as schematized in this brief introduction.

2. The evaluation of WTP and WTA.

A first family of SP methods is the one used in the estimation of WTP (or WTA) for changes in the provision of non-market goods. This category can be split into two separate subgroups: Contingent Valuation (CV) methods and Choice Modelling (CM).
CV is a stated preference method where respondents are asked about their maximum WTP for a predetermined increase or decrease in the consumption of a certain good.
The CM approach prefigures the design of a series of choice sets, each one containing usually two or more alternative scenarios. A scenario is a combination of several attributes, which are supposed to be relevant in determining individual choices; from each choice sets, respondents are asked to select the alternative they prefer. CM analysis can take different forms, based on the technique chosen, as summarised in table 1 (OECD, 2006).

Whatever the design of the questionnaire, the final goal of these studies consists in obtaining a perfect substitute for the unobserved RP. Hence, the optimal result is the one where the preferences stated by the respondents coincide with the behaviour they would follow if the hypothetical alternatives were true. Since real preferences are unknown, it is not possible to infer the quality of the analysis through a comparison between RP and SP.

Table 1. SP methods.

Method	Description
Contingent valuation	Respondents are directly asked about their monetary evaluations for a hypothetical scenario
Choice modelling:	
Choice experiments	Choose between two or more alternatives
Contingent ranking	Rank a series of alternatives
Contingent rating	Score alternatives on a scale of values
Paired comparisons	Score paired of scenarios on similar scale

However, economic theory provides some assumptions about consumers' choices. Moreover, it suggests how agents will behave when they are asked to state their preferences. For example, they could have incentives to misreport their behaviour.
Then, based on our theory, we can summarise *ex ante* all the sources of bias we expect will affect the SP study and all the assumptions about preferences which are supposed to hold. As a second step we can check, *ex post*, whether SP results are consistent with our *ex ante* hypothesis or not. If it is not the case and if we believe in our prior assumptions, inconsistent choices are not a good substitute for RP.

The possibility to check SP results is constrained by the design of the questionnaire (Foster and Mourato, 2002), whilst the theoretical assumptions depend on the approach we decide to follow. Mainstream economics, *i.e.* ordinal utility theory (Hicks and Allen, 1934), assume individuals to be rational and preferences to be consistent with this hypothesis. However, the so called *behavioural programme* (Thaler, 1988) rejects this assumption, showing that in many cases agents do not act in accordance with the paradigm ordinal utility theory.

2.1. SP reliability and ordinal utility theory.

In "A reconsideration of the theory of value" (1934) Hicks and Allen revised Alfred Marshall's theory of consumer behaviour on ordinalist lines, following Pareto demonstration of the immeasurability of utility (Pareto, 1909).

Some years after their contribution, Samuelson developed a new approach to consumers' behaviour, in an attempt to formulate a new economic theory, which may achieve the main results of ordinal utility theory, but being entirely based on observable phenomena and free from all unnecessarily restrictive conditions (such as, for example, the law of diminishing marginal rate of substitution, DMRS), dropping off, using Samuelson's words, last vestiges of the utility analysis. In his 1938's article, Samuelson gave the first description of the concept which ten years later, in 1948, he called "revealed preference" and formulated the postulate well-known as the Weak Axiom of Revealed Preferences (WARP).

Both ordinal utility theory and Samuelson's approach need some assumptions about the structure of preferences and the choice situation, in order to be able to derive preferences from observed behaviour.

Some properties regard the nature of consumers' preferences (Mas-Colell and Green, 1995). Let x, y, z be consumption bundles in the consumption set $X \in \mathbf{R}_{+}^{L}$. Then, consumer's preferences are defined as rational when they respect the assumption of:

a. completeness: for all $x, y \in X$ we have that $x \succsim y$ or $y \succsim x$ (or both);
b. transitivity: for all $x, y, z \in X$, if $x \succsim y$ and $y \succsim z$ then $x \succsim z$;
c. asymmetry: if $x \succ y$ it cannot hold the opposite, $y \succ x$;
d. non-satiation: for every $x \in X$ and every $\varepsilon > 0$, there is $y \in X$ such that $|y - x| < \varepsilon$ and $y \succ x$;
e. strict convexity: for every x, if $y \succsim x$, $z \succsim x$ and $y \neq z$ implies $\alpha y + (1-\alpha) z \succ x$ for all $\alpha \in (0,1)$.

Assumptions about individual preferences define the so called *homo economicus*, *i.e.* the abstraction used as a general model of motivation for economic agents.

In addiction to the previous five assumptions we need other hypotheses about the choice situation:

f. each good is perfectly divisible;
g. it is possible for a consumer to buy any combination of good he can afford;
h. consumers act as price-takers;
i. the price per unit of each good is the same regardless of the quantities purchased.

According to this theory individuals are always able to choose the market option they prefer. As a consequence, what could we expect from a SP analysis about individual preferences? As we said, the final goal of the researcher consists in possibly obtaining a perfect substitute of RP. Relative to the latter, SP methods can be biased by two factors:

1. individuals could falsify their own preferences in a way they find somehow beneficial;
2. choices may not respect the assumptions *a-e* listed above, as pointed out by several authors (Saelensminde, 2002).

Case 1: unfair statements. As we said, individuals may find convenient to misreport their choices. Economic theory provides some expectations about rational respondents' behaviour when they are asked to reveal their preferences.

In order to reach a Pareto-optimal provision for a public good, Samuelson Rule (Samuelson, 1954) requires that the total marginal benefit of the provision of another unit of the public good must equal the marginal cost of this extra unit. Since the total benefits are defined as the sum of individual ones, the policy-maker should observe the demand function of each individual.

The Lindhal equilibrium (Lindhal, 1958) provides a tax scheme which allows to reach a Pareto-efficient equilibrium through personalised payments based on individuals' marginal benefits. However, in many practical situations such mechanism fails when consumers are asked to reveal their true demands and may gain by making false statements of their preferences. The problem of preference revelation led to the definition of new mechanisms which make unprofitable any attempt of falsification.

The Clarke-Groves Mechanism (Clarke, 1971, Groves and Loeb, 1975) presents a contribution scheme for which truth telling is a dominant strategy. Basically, according to the model formalised by Groves, individuals are required to reveal the net benefits they get from the provision of a public good. The latter is provided only if total net benefits are positive. Then, each agent is subject to a transfer which is equal to the reported net benefits of other individuals. These side payments induce agents to tell the truth, since misreporting their preferences does not lead anymore to a gain in utility.
The Clarke Mechanism works similarly; the main difference concerns the transfer, which is applied only to pivotal agents, whose benefits change the social decision about the provision of the public good. Hence, compared with the Groves Mechanism the Clarke scheme lowers the cost of information revelation.

These mechanisms involve many practical problems (Rothkopf, 2007). In SP surveys, for instance, individuals are not subject to any form of transfer or side payment and their statements are at the mercy of dishonest incentives. In such a context the role of SP methods needs further investigation: if respondents are supposed to misreport their preferences, what is the benefit of asking them?

Another approach (Crawford and Sobel, 1981) concerns informative lobbying.
According to these models a lobbyist may have superior information about the policy environment than the policy-maker. By transmitting this information to the policy-maker he could positively contribute to the economy.
More formally, consider a model characterised by only two possible states of the world, θ_l and θ_h, where $\theta_l < \theta_h$. At the time the policy has to be chosen, the policy-maker is uncertain about the economic environment (θ). Each state of the world requires a different policy choice, respectively π_l and π_h if θ_l or θ_h are verified. Without additional information, the policy-maker would base his policy choice upon some prior beliefs about the possibility to observe alternative environments. However, if the policy is not correct for the state of the world which is realized, the society will incur in a welfare loss, according to the objective function:
$W(\pi,\theta) = -(\pi - \theta)^2$
If the policy-maker is uninformed about the environment and initially regards the two states as equally likely he will choose a policy based on the expected state of the world:

$$\pi^e = \frac{\theta_l + \theta_h}{2} \qquad (1)$$

Suppose the existence of a lobbyist who knows the state of the world. His preferences are defined by the function:
$U(\pi,\theta) = -(\pi - \theta - \Delta)^2$
that is, the ideal policy for the lobbyist exceeds the ideal policy of the policy-maker by an amount Δ (extent of disagreement) for either states of the world.
The lobbyist can only report either θ_l or θ_h and if he is trusted by the policy-maker the policy chosen will be respectively π_l or π_h.
If θ_h is the true state, the lobbyist does not have any incentive to misreport the information. If the state is θ_l the lobbyist has a potential incentive to lie because a truthful report, if trusted by the policy-maker, would lead to a policy level π_l, below the ideal policy of the lobbyist, $\pi_l + \Delta$.
Hence, the lobbyist may prefer to claim that the state is high, obtaining a policy π_h rather than π_l, if the extent of disagreement is sufficiently large:
$(\theta_l + \Delta) - \theta_l \leq \theta_h - (\theta_l + \Delta)$
which reduces to

$$\Delta \leq \frac{1}{2}(\theta_h - \theta_l)$$

If this inequality is not satisfied, when the state of the world is low the welfare loss will be equal to:

$$W_1(\pi, \theta) = -(\pi_h - \theta_l)^2$$

which is worse compared with the case in which the policy-maker decides not to ask the lobbyist and sets the policy equal to the expected value of the state of the world. Then, asking the lobbyist is not always the best choice, as it could be welfare worsening.

The information transmission problem in SP surveys can be analysed in a similar manner. A certain state of the world (*e.g.* the shadow price for a non-market good) occurs, but the policy-maker is still not able to observe it, while some individuals are aware of it. Then, the policy-maker directly asks agents to honestly reveal their superior information. In the SP exercise each respondent can be seen as a lobbyist, who represents his own interest rather than a certain group of people and still does not share the same preferences as the policy-maker. Moreover, the extent of disagreement differs among agents.

Consider for example an economy characterised by the policy-maker and two individuals with symmetric preferences. First agent's preferences are defined by:

$$U_1(\pi, \theta) = -(\pi - \theta - \Delta)^2$$

whilst the utility function of the second respondent can be written as:

$$U_2(\pi, \theta) = -(\pi - \theta + \Delta)^2$$

Again, the policy-maker does not know which state occurs between θ_l and θ_h. Then, he can set a policy based on the expected state of the world (1) or directly ask the two agents about their preferences and set the policy chosen equal to their mean:

$$\pi(\theta_i) = \frac{\sum \theta_i}{n}$$

where the subscript *i* indicates individuals 1 and 2 (here $n = 2$)

In this case, individuals' incentive to misreport their preferences is based both on the observed state of the world and on the expected statement of the other agent.

If the economic environment is θ_h, the first respondent will always tell the truth, while the second one will consider the following options. If he thinks the other individual is reporting θ_h, he will do the same only if

$$\Delta \leq \frac{1}{2}\left(\frac{\theta_h + \theta_h}{2} - \frac{\theta_h + \theta_l}{2} \right)$$

which reduces to [1]

$$\Delta \leq \frac{1}{4}(\theta_h - \theta_l)$$

while he will lie, reporting θ_l, if $\Delta > \frac{1}{4}(\theta_h - \theta_l)$.

Similarly, if he assumes the other agent is reporting θ_l, he will truthfully state θ_h only if $\Delta \leq \frac{3}{4}(\theta_h - \theta_l)$ and he will choose to lie if the latter inequality does not hold.

If the economic environment is θ_l, the same reasoning applies for the second respondent, whilst the first one will always give an honest statement.

Summing up, whichever the state of the world, if the extent of disagreement is sufficiently low ($\Delta \leq \frac{1}{4}(\theta_h - \theta_l)$) the policy-maker can expect both respondents to tell the truth. In this case the welfare loss will be equal to zero.

If the extent of disagreement is larger than $\frac{3}{4}(\theta_h - \theta_l)$ the welfare loss will be equal to:

$$W(\pi, \theta) = -\left(\frac{\theta_h - \theta_l}{2}\right)^2$$

which is the same result the policy-maker would obtain by setting the policy equal to the expected state of the world (1). If $\frac{1}{4}(\theta_h - \theta_l) < \Delta \leq \frac{3}{4}(\theta_h - \theta_l)$ the size of the welfare loss depends on the beliefs of each respondent about the opponent's choice. When the latter is equally likely, *i.e.* the probability attached to each statement (θ_l and θ_h) is the same, the respondent is indifferent between a fair answer and an unfair one. If it is the case the expected welfare loss will be equal to zero with probability $p = \frac{1}{2}$ and equal to $-\left(\frac{\theta_h - \theta_l}{2}\right)^2$ with probability $p = \frac{1}{2}$. Hence, whatever the state of the world and the size of the extent of disagreement, asking the agents cannot lead to a welfare loss lower than the one which derives by setting the policy equal to the expected economic environment.

Adding more respondents. Suppose now to add more lobbyists to our economy, keeping balanced the number of individuals whose preferences are above and below the state of the world. In other words, for each respondent with preferences equal to $(\theta + \Delta_i)$ there is another agent who would prefer a policy equal to $(\theta - \Delta_i)$. Note that, since the state of the world is defined by the mean of individual preferences, also the size of the extents of disagreement has to be balanced.

For instance, consider the case in which the economy is characterised by four individuals. Compared with the previous case the incentive to tell the truth reduces, as now the decision about the fairness of the statement is based on the behaviour followed by all other three agents.

Following the same reasoning as before, the policy-maker can expect all respondents to tell the truth if $\Delta_i \leq \frac{1}{8}(\theta_h - \theta_l)$. In this case the welfare loss will be equal to zero. On the

contrary, if the extent of disagreement is very large ($\Delta_i \geq \frac{7}{8}(\theta_h - \theta_l)$), half of agents (those with preferences $\theta - \Delta_i$ when the true state is θ_h and vice versa) will report the false environment, leading to a welfare loss still equal to the one reached by setting the policy chosen equal to the expected state of the world (1).
When the extent of disagreement is comprehended between these two levels, the decision about the statement is based on the beliefs each agent attaches to the opponents' choices. If the latter are equally likely, this time the best strategy is to tell the truth if $\Delta_i \leq \frac{3}{8}(\theta_h - \theta_l)$ and to misreport one own preferences if $\Delta_i \geq \frac{5}{8}(\theta_h - \theta_l)$, while between these two boundary levels each agent is indifferent between a fair or unfair answer. Again, asking individuals about their preferences leads, in the best case, to a welfare improvement and, if the extent of disagreement is very large, to a result which equals the one obtained by setting the policy at the environment's expected value. In general, keeping balanced the signs and the size of the extents of disagreement and increasing the number of lobbyists, we will expect a truthful statement by the whole sample if

$$\Delta_i \leq \frac{1}{2n}(\theta_h - \theta_l)$$

whilst half of the sample will report an unfair answer if

$$\Delta_i \geq \frac{(2n-1)}{2n}(\theta_h - \theta_l)$$

If the probability attached by each respondent to the opponents' statements is the same for θ_l and θ_h we will expect all individuals to tell the truth if:

$$\Delta_i \leq \frac{(n-1)}{2n}(\theta_h - \theta_l)$$

while half of the sample will misreport their preference if:

$$\Delta_i \geq \frac{(n+1)}{2n}(\theta_h - \theta_l)$$

Asymmetric preferences. The reasoning reported above holds if we assume preferences to be symmetric. In other words, the incentives to lie have to be perfectly balanced between individuals. However, this hypothesis sounds unrealistic in many empirical scenarios. For example, consider the case where the policy-maker is uncertain about whether or not to build a new high-speed railway. His decision is based on travellers' WTP for a reduction in travel time. Then, he decides to ask them about their valuation of time. However, in most cases travellers are worried about an increase in fares and have an incentive to lower their WTP $(\theta - \Delta_i)$. On the contrary, a minority group of wealthy respondents is enthusiastic about the project and their incentive to misreport preferences affects the results in the opposite direction $(\theta + \Delta_i)$.

Recall the first case and suppose to have three agents instead of two. The first two individuals' preferences are defined by:

$U_1(\pi,\theta) = -(\pi - \theta + \Delta)^2$

whilst the utility function of the third respondent is:

$U_2(\pi,\theta) = -(\pi - \theta - \Delta)^2$

Suppose θ_h to be the true state of the world. Then, the third individual will always tell the truth, while the others will base their decision on the expected behaviour of the opponents. They will always truthfully reveal their preferences if $\Delta_i \leq \frac{1}{6}(\theta_h - \theta_l)$. Vice versa, they will lie when $\Delta_i \geq \frac{5}{6}(\theta_h - \theta_l)$. Between these two thresholds their behaviour will depend on the beliefs attached to the opponents' statements. If they do not have any particular presumption about them, they will correctly report the state of the world if $\Delta_i \leq \frac{1}{2}(\theta_h - \theta_l)$ and they will lie if $\Delta_i > \frac{1}{2}(\theta_h - \theta_l)$.

When θ_l is the true environment, the first two agents will always tell the true, whilst the third one will behave according to the same incentives depicted for his opponents.

If we consider the role of the policy-maker, he can set the policy chosen to the expected value of the state of the world (1) or decide to ask the individuals about their preferences. If the true state is θ_l the welfare loss will be at most equal to

$$W_l(\pi,\theta) = -\left(\frac{\theta_h - \theta_l}{3}\right)^2$$

which is better compared with the welfare loss which derives by setting the policy according to (1):

$$W_1(\pi,\theta) = -\left(\frac{\theta_h - \theta_l}{2}\right)^2$$

However, if the true state of the world is θ_h, asking the individuals could lead to a welfare worsening, since if first two agents' extent of disagreement is large enough the loss will be equal to

$$W_h(\pi,\theta) = -\frac{2}{3}(\theta_l - \theta_h)^2$$

instead of

$$W_1(\pi,\theta) = -\frac{1}{2}(\theta_l - \theta_h)^2$$

Compared with the case characterised by symmetric preferences, this time the profit from asking individuals depends on the size of the extent of disagreements. Compared to the results from previous literature on this topic (Hindricks and Myles, 2006), the presence of more than one informed agent does not necessarily help the policy-maker in determining the best policy to implement.

As we have seen, economic theory suggests that, without any monetary mechanism attached to the process of revelation of preferences, dominant truth-revealing strategies do not hold. Agents' statements could be affected by incentives to misreport true preferences. This source of bias would distort SP results compared with unobserved RP. However, in some cases the policy-maker finds convenient to ask agents about their preferences because, by doing so, the expected welfare loss is not lower than the one he would obtain by setting the policy chosen at random.
Whatever the case, if individuals lie according to assumptions *a-e* listed above, the researcher will not be able to distinguish between fair and unfair statements.

Case 2: inconsistent choices. Some authors (Foster and Mourato, 2002) pointed out how, in many SP studies, a significant share of answers is not consistent with the predictions of economic theory. Hence, they suggest to reduce the sample through the exclusion of all illogical answers. Since the evidence shows that the inclusion or the exclusion of test failures has a significant effect on WTP estimations, the problem is relevant.

The possibility to test theoretical assumptions is constrained by the structure of the SP study. Forster and Mourato (2002) apply their reasoning to a *contingent ranking* experiment (see table 1). Salesminde (2002) verifies logical consistency by using a *choice experiment* survey.

Then, reliability of SP results can be checked only based on the hypotheses which define the *homo-economics* portrait. According to this approach, reliability is defined in terms of logical consistency to assumptions *a-e* summarised above.

2.2. SP reliability and the behavioural economics programme.

A branch of economic analysis, the so called behavioural programme, investigates what happens in markets in which some of agents display human limitations and complications (Mullainathan and Thaler 2000).

According to this approach (Kahneman, 1997), two core meanings of utility can be distinguished. *Decision utility* is inferred from choices and used to explain them. It equates happiness with choices *per se*, since utility is defined in terms of choice itself. Under this approach (Kahneman and Thaler 2005) what matters is the question of whether preferences are consistent with each other and with the axioms of rational choice, such as those listed above. *Experienced utility*, in contrast, is hedonic quality, as in Bentham's usage. It is based on desires, as it equates happiness with desire fulfilment. According to Bentham (1789), by the principle of utility *is meant that principle which approves (...) of every action whatsoever, according to the tendency which it appears to have to augment (...) the happiness of the party whose interest is in question*.

Relative to the ordinal utility theory, two new terms are introduced: desire and happiness. The former is the sole element which drives human's behaviour, while the

latter constitutes the target of every human's act. Since both concepts are non observational, if we want to rationalize individuals' behaviour we have to refer to some observable substitutes. Ordinal utility theory operated this substitution by letting coincide desires with choices and happiness with utility. Moreover, it is assumed that consumers' preferences can be rationalized according to some rules (assumptions *a-e* above) and that agents never fail to choose the market option which maximizes their utility. The behavioural economics programme criticized the *decision utility* approach since, by divorcing utility from desire altogether, it merely becomes a *theory (...), of the numerical representability of choice* (Sen, 1981). The criticism is based on two fundamental issues.

First of all, choices and utility do not always work well as substitutes for desires and happiness. If choice sets and utility levels are defined on the base of observable entities such as prices and quantities, many times the concepts of desire and happiness involve other, non observational ingredients such as moral and ethical issues. For this reason, consumers' behaviour cannot be rationalized based on those assumptions which characterize the *homo economicus* abstraction. Observed market behaviour shows how people do not always choose according to those rules they are supposed to follow, at least for those cases in which the maximisation of their utility does not coincide with the maximisation of their desires.

Consider the following example, extracted from the large literature existing on this subject (Bowles, 2004). In an ultimatum game two players have to split 100 $. The first player makes a proposal to the second one, who can accept or turn it down. Consider now a subject who rejects an offer of splitting 100 $ by (92$; 8$), after realizing that the first player could have proposed a more equitable partition. Choices of this kind are often observed (Rabin 2002), even if they are not consistent with predictions of economic theory and in particular with the assumption of non satiation (assumption *d* reported above). A preference for the allocation (0$; 0$), rather than for a more profitable one as (92$; 8$) looks irrational under a neoclassical point of view but fully motivated from a human perspective.
Another example, concerning again departures from the self interest assumption, is the following. Player 1 is asked to choose between two different allocations for two other individuals.

Table 2.

Player 1 chooses			
	Money for Player 2	**Money for Player 3**	*Approximate findings*
Option 1	7.50 $	3.50 $	*50%*
Option 2	4.00 $	4.00 $	*50%*

Empirical evidence, reported above in table 2 (Robin and Charmess 2000), shows how half of the respondents prefers the solution that maximizes the total outcome (option 1), while the others are more concerned about the situation of the poorest individual.
Now let change the rule of the game: it is Player 3 who has to choose between the two allocations. Results, reported in table 3, are quite similar compared with the previous case excepted for a slightly stronger preference for the second option, which is more advantageous for Player 3 himself.

Table 3.

Player 3 chooses			
	Money for Player 2	**Money for Player 3**	*Approximate findings*
Option 1	7.50 $	3.50$	*40%*
Option 2	4.00 $	4.00 $	*60%*

Suppose now that, again, Player 3 has to choose as in the previous case but, this time, he chooses after that Player 2 has created this scenario by rejecting a third alternative (5.50$; 5.50$). The decision of Player 2 to remove the latter allocation in favour of trying to get Player 3 to choose the first option is an unfair behaviour, as it involves a small increase in total surplus while leading to an unequal distribution of money.

Table 4.

Player 3 chooses after Player 2 rejected (5,50; 5,50)			
	Money for Player 2	**Money for Player 3**	*Approximate findings*
Option 1	7.50 $	3.50$	*10%*
Option 2	4.00 $	4.00 $	*90%*

This time (table 4) Player 3 chooses quite differently: the strong majority of respondents are not willing to sacrifice 0.50 $ in favour of a better allocation for Player 2. It has to be noted how, in the last two cases, Player 3 was facing the same problem. Even if the alternatives were identical, his choice changed in response of previous behaviour of other players. In this case the assumption of symmetry (*c* in the list above) does not hold: some individuals who preferred option 1 over option 2 in the second exercise inverted their preferences in the last experiment.

Another objection moved to the traditional framework regards the evidence that people care a lot about changes, and not only in absolute level of consumption, wealth or whatever. For example, the event of becoming wealthy, not only being wealthy, can

often be a major source of satisfaction. Rabin (2002) notes that a crucial role in explaining preferences is played by loss aversion, since individuals often value losses much more than commensurate gains. The so called *endowment effect* represents, according to Knetsch, Tang and Thaler (1998) one of the most robust findings of the psychology of decision making [2]. The utility which derives from the current consumption not only depends on its level but also on how that level compares to what we are used to. Camerer (1995) suggests how reference-dependence can be thought of simply in terms of a new assumption about preferences: letting c be a vector of the levels of, let's say, consumption of goods, and r be a vector of reference levels in the same dimension, incorporating reference dependence into the utility theory involves merely a switch from $U(c)$ to $U(c,r)$.

At last, a prediction of economic theory that is frequently rejected by experimental evidence concerns preferences over time. Individuals prefer to experience pleasant things soon and to delay unpleasant things until later. As pointed out by Rabin (2002) economists traditionally model such tastes by assuming that people discount streams of utility over time exponentially. But this functional form generates time consistent preferences, i.e. the reference between any two intertemporal tradeoffs in monetary well-being is the same no matter when asked. Behavioural evidence shows, in contrast, how people *exhibit present-biased preferences: they are more averse to delaying today's gratification until tomorrow than they are averse to delaying the same gratification from 90 days to 91 days from now* Rabin (2002). Many psychological experiments support present-biased preferences and some researchers (Gollier 2003) tried to model such preferences through alternative discount functions.

We have presented three cases in which choices do not respect traditional assumptions of economic theory even if they cannot said to be irrational: people' preferences are not always driven by self interest, agents care not just about final outcomes but also how they arrived there and how their situation changed, their preference are present-biased. The violations are due to the incompleteness of our hypothesis, which fail to consider all the elements that do matter in consumers' behaviour.

The second criticism moved to mainstream economics concerns the cases in which agents fail to choose the market option which maximizes their utility (or their happiness) because of bounded rationality. As Thaler (1988) refers, in the traditional framework economic agents are supposed to be able to make accurate, or at least unbiased forecasts of the hedonic outcomes of potential choices. Systematic errors in utility forecasting have been detected in several cases. A good example is provided by Nisbett and Kanouse (1968) about hungry shoppers: they showed that shoppers who are hungry tend to buy food as they expected to remain permanently famished, but shoppers who are given something to eat before entering the supermarket are more likely to restrict their shopping to the items on their list (Gilbert, Gill and Wilson, 1998). As pointed out by Kahneman and Thaler (2005), if the current state of hunger induces the shopper to buy

an overly large dinner portion for consumption later in the week, then he has made a forecasting error which has lead to a bad choice.

Kahneman (2003) provides a detailed literature review about bounded rationality. Basically, two generic modes of thinking and deciding are identified. The first one (System 1) refers to what is usually defined as "intuition", while the second one (System 2) corresponds to the everyday concept of "reasoning". If the operations of System 1 *are fast, automatic, effortless, associative and often emotionally charged* (…) System 2 is *slower, serial, effortful and deliberately controlled* (Kahneman, 2003). When asked to take decisions or to express a judgement, individuals base their behaviour on both impressions and tendencies generated by System 1 and on the monitoring and corrective functions of System 2. The performance connected with each system is affected by several factors. As reported by Kahneman, the ability to avoid errors of intuitive judgement is impaired by time pressure, by concurrent involvement in a different cognitive task, by performing the task in the evening for "morning people" and vice versa, by being in a good mood. The facility of System 2 is positively correlated with intelligence, with the so called "need for cognition" (a psychological finding which characterise those people who find thinking fun) and with exposure to statistical thinking.

The *behavioural programme* provided empirical demonstration of some weakness of ordinal utility theory. Assumptions *a-e* listed above are not always verified and then, according to the behavioural approach, they cannot be used to discriminate between reliable and unreliable behaviours. Hence, keeping in mind that the final goal of the SP methods consists in obtaining a perfect substitute for real and unobserved preferences, which sources of bias could we expect to have an influence on our results?
Again, SP methods can be biased by two factors:

1. individuals could falsify their own preferences in a way they find somehow beneficial;
2. agents may fail to report the preferred option due to bounded rationality.

Case 1: unfair statements. As in the previous case, individuals could still have an incentive to influence the policy chosen by the policy-maker. However, from a behavioural perspective it is not clear how misreported preferences will affect social welfare. Individuals' choices cannot be rationalised as in the previous paragraph: rather then basing their statements on opponents' expected behaviour, agents decide to lie on the ground of a mixture of moral sentiments and biased (or unbiased) beliefs. For instance, Bonsall (1985) refers to *justification bias* for those cases in which individuals deliberately give biased choices simply because they want to look in a better light with the interviewer. Whatever the purpose of the unfair statements, it is still not possible to discern, *ex-post*, between fair and unfair statements.

Case 2: unintentional misreported preferences. In some situations individuals may fail to correctly report their preferences even without being motivated by dishonest intents.

Here statements' accuracy cannot be interpreted anymore in terms of logical consistency with respect to assumptions *a-e*. The first group of objections moved to mainstream economics showed how the *homo-economicus* abstraction fails to keep into account elements which do matter in the definition of individual utility patterns.
In the SP exercise agents are asked to choose, rank or score different market options, each one characterised by a certain level of the relevant attributes. Then, answering the survey implies a mental effort, and the intensity of this effort is related to both survey's design and individual skills.

Many studies demonstrate how survey's design affects agents' statements. Intuitively, the greater the number of alternative scenarios, the more intense the cognitive effort. Boredom effects and response fatigue have been confirmed by some researchers (McFadden, 1986). Often agents try to minimize their mental strain by following rules of thumb or anchoring to earlier tasks (Bates, 1988). The problem can be analysed recalling the classification suggested by Kahneman (2003). When asked to choose between two alternative scenarios or to rank some market options, agents deal with a logical problem and they use System 2 in order to understand the rules of the game and to think about their answer. However, if the choice problem is reiterated (the interviewee is asked to choose between a sequence of couples of alternative options) the respondent is likely to base his statements on rules of thumb rather than on effortful reasonings. For instance, he could take as reference a certain level of an attribute and accept (or refuse) all the options characterised by a higher (lower) quantity of that particular attribute. In other words, agents shift from System 2 to System 1 since they want to reduce the fatigue connected to the SP exercise. In this way they may fail to choose the preferred option because they do not pay sufficiently attention to the alternative options.

Independently of survey's design, also individual skills affect agents' capability to correctly report their preferences. As noticed by Kahneman (2003) individuals who are used to deal with logical problems are more likely to properly solve the SP exercise.
Individuals characterised by high logical skills are also more likely to elaborate efficient rules of thumb.

Compared with the case presented in the previous paragraph, the behavioural programme does not provide a set of assumptions about consumers' behaviour. Hence, it is not clear how the consistency of SP results can be checked and in which cases they should be rejected from the survey. The next section is devoted to further discussion about this issue.

2.3. SP reliability: discussion.

We analysed as a first category of SP those analysis in which agents are asked to choose among different and alternative scenarios, where each option is characterised by different levels of some attributes of interest for the policy maker. As in a market-choice

situation, individuals evaluate and compare the alternatives and choose the preferred one. In general, the closer the SP are to the unobserved RP, the better is the quality of our survey and, consequently, the more reliable its results are. Since, as we said, RP are not observable, in principle it is impossible to discriminate between reliable and unreliable stated preferences. The only possible way of proceeding is to make some assumptions about RP and to exclude from the surveys all those statements which are not consistent with our hypothesis.

In the previous sections we described two different approaches to the same problem: the first one based on mainstream economic theory and the other one in accordance with the behavioural economics programme. These methods provide different sets of assumptions about consumers' behaviour. On the ground of these hypotheses SP results can be checked for logical consistency (LC). However, this procedure presents some problematic aspects which concern both approaches.

Firstly, consider the mainstream economics viewpoint.
Some authors (Foster and Mourato, 2002) showed how, in many SP analyses, some agents fail to report consistent preferences compared to the predictions of economic theory. Then, following this approach we are adopting a *decision utility* perspective: all those choices which are logical consistent with our axioms are reliable, the others are not. On the one hand this method sounds correct. Through SP analysis we are, by definition, trying to derive preferences from observed (in this case stated) choices. Researchers use statements in order to estimate a utility function whose elements are the attributes included in the SP exercise. As this is nothing but a *decision utility* approach, there is no reason for which we should not analyse the results of such a study from the same perspective.

On the other hand this procedure seems somehow contradictory. Since consumer theory is supposed to explain consumer's behaviour, why should we observe cases which violate our theoretical assumptions? Moreover, if we observe such inconsistent preferences and we decide to exclude them from the sample, we are admitting that some agents behaved irrationally, which turns out to be a rejection of our theory.

One could object that some respondents are not interested in answering the survey. When asked to give their contribution to the SP exercise they are facing two costs. The first one (c_1) is the cost of answering the questionnaire, which depends on the mental efforts and on the revelation of private information. The second one (c_2) is the cost for refusing to participate to the SP exercise, and it is connected to the bad impression the respondent would make with the interviewer if he refuses to help him. If c_1 and c_2 are sufficiently high, the best strategy for the respondent is to participate to the survey by giving answers at random, in order to minimise the mental effort and please the interviewer. Then, with a certain probability he will not pass the LC test but, in a real context, he would act according to assumptions *a-e* listed in the previous section.

Following this reasoning the attention shifts from the reliability of preferences to the validity of our hypothesis: if the latter holds, then a certain set of properties can be used as an instrument of detection of unacceptable preferences.

In a sense, ordinal utility theory can be viewed as a sort of progenitor of the behavioural programme. According to Hicks and Allen (1934), who revised Alfred Marshall's theory of consumer behaviour on ordinal lines, all concepts must correspond to observational phenomena, as Pareto provided a *positive demonstration that the facts of observable conduct make a scale of preferences capable of theoretical construction (...) but they do not enable us to proceed from the scale of preference to a particular utility function.* The first victim of this reasoning is the concept of cardinal utility, and the rule which drives consumers' choices becomes the principle of Diminishing Marginal Rate of Substitution (DMRS). The latter, that allows indifference curves to be convex to the axes and, consequently, enables the equilibrium to be stable, *is not a mere translation; it is a positive change in the foundation of the theory, and requires a very definite justification* (Hicks 1939, p. 21). A first justification is based on empirical evidence. People buy some quantities of commodities and reject other market options, then it follows that the principle of DMRS must sometimes be true. Hicks recognizes that this explanation cannot give general validity to the new theory and therefore tries to suggest a more persuasive argument. He states that when market conditions change, the consumer moves from one equilibrium to another. At each of these positions the condition of DMRS must hold, otherwise he could not take up such a position at all. In order to proceed from this to the law of DMRS it is necessary to assume that the condition holds at all intermediate points between two indifference curves and, in other words, that there no kinks between the two positions of equilibrium. Hicks concludes that *being the simplest assumption possible, it is as good assumption to start with; and in fact its accordance with experience seems definitely good* (Hicks 1939, p.24).

The foundation of these arguments has been questioned by several authors. Wong (1978, p.38) suggests that *Hicks, by confusing two separate issues, makes the issue of truth subsidiary to that of theoretical necessity*. In his explanation, based on observed consumers' behaviour, Hicks implicitly assumes that individuals always maximize their utility, under the constraints represented by the current system of prices and their available income. This justification turns into a circular argument rather than in a satisfactory proof for the universal validity of the principle of DMRS. As pointed out by Robinson (1962) *utility is the quality in commodities that makes individuals want to buy them, the fact that individuals want to buy commodities show that they have utility.* Then, the structure of preferences cannot be justified from the evidence that people do buy, because we are already assuming that people buy in accordance to those preferences that we are trying to infer from their behaviour.

Similar objections to Hicks' arguments have been moved from Paul Samuelson in 1938. For Samuelson, even if Hicks and Allen worked in this direction, ordinal utility theory failed to become an observational theory. In particular, an objection was addressed to

the principle of DMRS and to Hicks's proof. Samuelson (1938) asks: *why should one believe in the decreasing rate of marginal substitution*[3]*, except in so far as it leads to the type of demand functions in the market which seem plausible?*

Based on this objection, Samuelson's objective, in 1938, consists in developing a new economic theory, which may achieve the main results of ordinal utility theory, but being entirely based on observable phenomena and free from all unnecessary restrictive conditions, dropping off, using Samuelson's words, last vestiges of the utility analysis. In his 1938's article, Samuelson gave the first description of the concept that ten years later, in 1948, he called "revealed preference" and formulated the postulate well-known as the Weak Axiom of Revealed Preferences (WARP). He states that *if an individual selects batch one over batch two, he does not at the same time select two over one* (Samuelson, 1938). More formally, the commodity bundle x is revealed preferred to the commodity bundle y (written xSy) if for some competitive budget x is chosen when y is affordable. According to the weak axiom of revealed preferences, if xSy we cannot have ySx.

It is important to stress how Samuelson always used, in 1938, the term "select" instead of "prefer", following his attempt to build a theory based on individuals' behaviour, without any reference to the concept of utility. However, as pointed out by Wong (1978), Samuelson theory fails, until now, to be an explanation of consumer behaviour. While Hicks and Allen theory explains a consumer's behaviour in terms of his preferences and his given material constraints, it is impossible to infer from Samuelson theory the reasons why one bundle was bought and all the others bundles were not.

This issue is faced by Samuelson in 1948's "Consumption Theory in Terms of Revealed Preference". Here, he states that *if enough judiciously selected price-quantity situations are available for two goods, we may define a locus which is the precise equivalent of the conventional indifference curve*. In this way, *by comparing the costs of difference combinations of goods at different relative price situations, we can infer whether a given batch of goods is preferred to another one; the individual guinea pig, by his market behaviour, reveals his preference pattern, if there is such consistent pattern*.

This seems in strong contradiction with the previous article and, as stated by Wong (1978) it is *highly questionable that the present aim of constructing an indifference map is a legitimate extension of a theory in which the use of utility or any other non-observational concept is diligently avoided*. The term "preference" has now a central role, while in the previous article it has never been used. Then, the purpose of Samuelson's analysis is, in 1948, to drawn consistent preferences from observed consistent behaviour, where the consistency is defined according to the WARP.

According to Wong (1978) this inference requires two assumptions. First of all we have got to assume that preferences do exist. Then, we need an assumption about consumers,

who are supposed to act in accordance with their preferences subject to material constraints. This is almost equivalent to assume the truth of ordinal utility theory and of its set of assumptions, which are not open to verification since, in order to prove their validity, we should be able to consider every possible combination of goods.. Consequently, Samuelson theory suffers of the same limitation, i.e. the circularity of the argument about preferences. Instead of being a new theory of consumer behaviour, it becomes an empirical verification of ordinal utility theory and a means of revealing consistent preferences.

More recently, as we have seen in the previous paragraph, behavioural economics provided several examples in which mainstream economics fails to correctly predict consumers' behaviour. Then, if sometimes individuals maximize their preferences not in accordance with the tenets of mainstream economics why should we base our test for logical consistency on a set of properties which do not have universal validity?
Why individuals should be supposed to act more rationally in surveys than in reality?

Adopting such an approach could lead to one of the following undesired situations listed by Wong[4] (1978):

1. behaviour which violates assumptions *a-e* does not imply that the consumer's preferences are inconsistent;
2. behaviour which satisfies the assumptions *a-e* not imply that the consumer's preferences are consistent;
3. inconsistent preferences may not necessarily detected through the use of a logical consistency (LC) test based on assumptions *a-e*.

In the first case, behaviour which violates traditional assumptions does not imply that the consumer's preferences are inconsistent. Consider for instance those individuals who, in the ultimatum game reported above, accept an allocation of (0$; 0$) rejecting a more convenient one of (92$; 8$). Excluding such choices would be incorrect as, in this case, agents fully understood the problem, evaluated the alternatives and correctly reported their preferences. These three conditions assure reliability of results, no matter whether they respect the *homo economicus* psychological profile or not.

On the other hand, and here we are at the second undesired situation, behaviour which satisfies our axioms does not imply that the consumer's preferences are consistent. Let's think at the example about the hungry shopper. He fails to maximise his utility because of his current mental state. Analogously, while answering a survey, agents' predictions about future utility could be biased by their current mental and emotional state. Maybe they will choose alternatives consistently one with the others and they would pass a logical consistency test, even if they selected a sub-optimal allocation.

At this point an objection could be moved: where is the difference between a hungry shopper and an angry player, rejecting an allocation of (92$; 8$) in favour of another one of (0$; 0$)? The latter is not choosing optimally since, at t_0, the desire of revenge on

his opponent drives his preference toward an allocation that at t_1 he will regret. Hence, a consistency test that allows classifying such a choice as irrational works right. In principle this objection seems reasonable, but it does not keep into account a fundamental difference between the two cases. In the first one, the hungry shopper can report his (biased) preferences through choices which are fully consistent with the axioms of rational choice. If it is the case, all consistency tests based on those axioms will catalogue such behaviour as rational, even if it fails to maximise utility. In the second case the angry player will maybe regret his choice, but maybe not. We can suppose that it depends on the temporal distance between t_0 and t_1, but we can also assume that his sense of justice simply cannot let him accept a compromise, no matter how much the time passed. However, testing logical consistency would lead, in this case, to a verdict of irrationality and to a consequent exclusion of these choices from the "clean" sample, even if the agent correctly anticipated his utility. Since it is not possible to check for dynamic consistency, as in SP we observe the choice in t_0 but we do not have any feedback about experienced utility in t_1, the exploitation of LC tests based on traditional assumptions about rationality involves two problematic issues: not only the exclusion of optimal choices but also the inclusion of sub-optimal ones.

Even if the behavioural approach seems more appropriate to describe individuals' behaviour, a problematic issue arises when we try to check the reliability of our results. The final goal of the researcher is to obtain, through the SP exercise, choices consistent with real preferences. Since in real markets we observe some situations where agents choose irrationally, how can we discriminate between reliable and unreliable statements? This issue can be analysed by considering separately the two groups of objection moved by the behavioural programme.

First of all, we saw that assumptions which define the *homo-economicus* abstraction fail to have empirical verification, since sometimes they miss to keep into account factors which matter in driving agents' behaviour. Recalling the example about the ultimatum game, in the third situation the player seems to choose irrationally, as he refuses a more convenient and available option. However if we observe the previous, unfair choice of his opponent, this behaviour does not seem irrational anymore. Simply (Sen, 1980), individuals' utility is driven by both quantities (of money, commodities, etc.) and unobservable values (fairness, solidarity, etc.). In this case agents act rationally, but rationality cannot be defined in the restrictive way suggested by mainstream economics.

In a SP context this kind of problem does not occur. Fictitious market scenarios provide a complete description of the economic environment. Utility and happiness coincide: since the alternatives are in general fully represented by their attributes, there is no reason to make allowance for unobservables, such as moral issues or ethical considerations.

The second group of objections concerns bounded rationality: sometimes agents fail to choose the preferred option because of their logical limitations. Hence, when we observe irrational choices in the SP exercise, how could we pretend to exclude them, if also in the real world individuals may behave in the same way? Moreover, how can we detect irrational choices if the assumptions on which mainstream economics bases the test for LC do not have universal validity?

The first question can be answered by comparing the context of the survey study with real markets. In SP analysis respondents are usually asked to compare and evaluate a large number of alternatives, while in real markets the available options are usually a few. The mental effort connected with the participation to the SP exercise is higher compared with the intellectual strain generally experienced by individuals. Then, we can assume that those agents who report irrational preferences in the survey would not act in the same way in a real market context.

But how to detect irrational choices? Recall again the ultimatum game. Let's say that this time the player who moves second can choose among three allocations: A(0$; 0$), B(92$; 8$), C(92$; 9$). Again, if he chooses option A because he wants to punish the unfair behaviour of his opponent, his preference cannot be defined as inconsistent. But the same reasoning does not hold when he opts for B since, given that the choice is not caused by the wish to damage the other player, this alternative is strictly dominated by C. Similarly, if we admit that preferences are reference-dependent, a consistency test can be carried out only among those choices which result from comparable scenarios.

In general, logical consistency can be checked only by taking into account and keeping constant all the parameters that can influence the decision process of individuals. This is the case of a SP exercise where, as we said, the alternatives are in general fully represented by their attributes. We can think at this issue by referring again at the questions suggested by Kahneman and Thaler (2005) and cited above. From a *decision utility* perspective, the relevant question is the following: are preferences consistent with each other and with the axioms of rational choice? A positive answer assures reliability. But this result is misleading, since consistency with the axioms of rational choice does not always guarantee neither a reliable substitute for RP nor utility maximization.

Then, from an *experienced utility* point of view, the issue of reliability could be addressed by adding a specification to the previous question: are preferences consistent with each other and with the axioms of rational choice, when the consistency with those rules is relevant, beyond any doubt, in explaining whether people choose the options they will most enjoy?

3. Surveys about attitudes and perceptions.

In this paragraph we analyse another family of surveys. While in the previous case individuals were asked to choose the preferred option among different scenarios, and the reiteration of the decision process allowed checking for logical consistency, here agents are required to reveal their attitudes and perceptions.

This kind of surveys can be divided into two groups:

1. the first group includes all those situations where agents are required to reveal their actual behaviour, which cannot be observed by the researcher. For instance, respondents may be asked about the number of hours spent watching TV, about their political views, their income, etc.;
2. the second group is composed by the surveys in which agents are asked to express their feelings or perceptions about a certain aspect of their life. For example, they may be required about their satisfaction with life or about their perceived health status.

In these cases there is no way of checking for logical consistency, since all choices are equally rational and, by choosing only once, people cannot contradict themselves.
Also the distinction between the mainstream economic and the behavioural approach does not hold anymore, since standard economic theory often rejects this kind of surveys, which are supposed to be unscientific and unreliable since their results are not objectively observable. Again, we are back to the conflict between *decision* and *experienced utility*. Followers of the traditional approach are persuaded that utility can be inferred directly from choices, while behavioural economists suggest that happiness is not only a matter of quantities (of commodities consumed, of money owned, etc.) but also depends on other, more qualitative, issues, undetectable through classical RP studies. How can be explained, for instance, the paradox that in several countries since Wold War II real income has drastically risen but the perception of subjective well-being of the population has not increased or has even fallen slightly (Frey and Stutzer 2001, Rabin 2002)?

In the same way as for the first family of surveys the issue of reliability is relevant in order to fully understand the usefulness and limitations of such analysis.
The main peculiarity which distinguishes the first group of surveys from the second one is the existence of a true and objective answer. For instance, when asked about the time spent watching TV agents may lie, because they find more appropriate to report a lower number of hours compared with the real one. However, if the researcher were able to observe agents' behaviour, he would know the correct answer, in this case the number of hours individuals truly spend watching TV.

Some researchers demonstrated the existence, also for this kind of surveys, of what we defined in the previous section as *justification bias*. Respondents may misreport their actual behaviour since they may avoid looking bad in front of the interviewer (Bertrand

and Mullainathan, 2001). Agents could also misreport their behaviour unintentionally, due to cognitive factors such as question wording, ordering of questions, etc[5]. Bertrand and Mullainathan (2001) report an example still concerning the time spent watching TV. A sample of German respondents was been asked about how many hours of TV they watch per day. A first group of individuals have been presented with a questionnaire with ten options, starting from "less than thirty minutes" and then proceeding in half an hour increments up till "more than four hours and a half". The second group of respondents had to answer a questionnaire where the lowest option was "two hours and a half". In the first sample, only 16 respondents out of 100 declared to watch more than two hours and a half of TV per day, whilst in the second group the same share was about 32 per cent. This evidence is due to the fact that respondents appear to infer "normal" TV viewing from the scale of the questionnaire.
Whichever the source of bias, it is impossible to evaluate the reliability of answers through an *ex-post* analysis. In this particular case attitudes are objective and known only by respondents. Then, the only method that could increase the reliability of such surveys works *ex-ante,* and consists in clarifying which cognitive factors can increase the bias of results and improving the quality of questionnaires' architecture (Sunstein and Thaler, 2008).

In the second subgroup of surveys, where agents are required to express their perceptions and feelings, the problem of reliability is more complex.
This time there is not any objective answer to the survey questions. Consider for example a researcher interested in investigating consumers' satisfaction about public transport. Suppose he asks two individuals who shared the same travel experience, and both of them honestly report their opinion. The first one reports a low level of satisfaction (say two out of ten), whilst the other is fairly satisfied (he scores seven). What can be learnt from such a result? With respect to the previous case, this time attitudes are not objective, as for the number of hours spent watching TV. Here perceptions are based on personal perceptions and evaluations which depend on several unobservable and subjective factors.

Some examples come from the health economic literature. McFadden *et al.* (2005) report the case of a survey conducted in France and Denmark about self-rated health status. Evidence shows that 62 percent of Danish men defined "excellent" their health status, while in France the same answer occurred only for 14 respondents out of 100. Since, according to health statistics, French men have a life expectancy two years greater than Danish citizens, the result of this subjective survey seems in contradiction with objective evidence. Hence, in this kind of studies the issue of reliability involves the connection between subjective perceptions and objective values. The topic is relevant since often policy-makers and politicians are interested in collecting public opinion surveys and they are strongly motivated to maximise individuals' satisfaction.
If the latter properly represents objective and unobserved values (such as in the examples above, the quality of public transport and agents' true health status) then surveys could constitute a useful support to public management and policy choice.

Recalling the example provided by McFadden (2005), one could object that correctly evaluating your own health status is much more difficult compared with reporting your satisfaction with public transport. In the first case the evaluation is based on personal feelings and individuals are forced to found their judgements on poor information, while when asked to rank their satisfaction with transport, they are perfectly aware of the price they paid, of coaches' cleanliness and of all other relevant features of the service. Consequently, the relevance of our question, i.e. whether or not satisfaction and objective quality coincide is related to the quantity of information on which individuals base their judgements[6].

Suppose we want to investigate satisfaction with healthcare services provided by the public sector in a certain country[7]. Imagine patients to be very satisfied (as they usually are), even if we observe that objective indicators of quality do not support their enthusiasm. For example, physicians may prescribe more medicines than needed or hospitalize patients when it is not strictly required. Therefore, individuals receive bad assistance but they do not have enough information (here, medical skills) to realize it. In this case the issue concerning the relation between satisfaction and quality is strongly relevant and the two concepts probably do not coincide. Suppose now to replicate the same experiment about public transport and to get the same result: respondents are very satisfied with railways even if we observe low objective quality, for example coaches are often in delays. The only conclusion we can infer from this evidence is that people have weak preferences for time savings, but we would not worry about it as in the previous example.

In general, service quality can be defined by the set of features which drives individuals' preferences. In the extreme (and unrealistic) case, if consumers are able to observe all these characteristics and to base their choices and evaluations on them, then quality and satisfaction coincide.

However, this reasoning does not hold anymore in a comparison between different groups of individuals asked to judge different items. Suppose to find evidence similar to the one showed by McFadden (2005) about public transport: Danish travellers are more satisfied than French ones. For instance, the former attribute on average a score of nine out of ten to transport service, while the latter give a rating of seven out of ten. From this result we could make two assumptions. The first one is about the quality of transport service in each country. As in the previous case, if we think that consumers' satisfaction properly describes objective quality, we can infer useful information from our survey. The second assumption concerns the information about the relative quality of one transport service compared with the other. In this case, we shall assume that transport services work better in Denmark than in France. But this interpretation does not keep into account the fact that we have two distinct items (public transport in Denmark and France) judged by two different groups of respondents (Danish and French).

In this context, the issue of reliability involves not only the concept of objectivity with respect to "real" quality, but also the problem of comparability between different groups of individuals. For example, due to some social norms or cultural factors, French could be more severe in judging public services than Danish, and this effect could partially explain the different pattern of satisfactions emphasized by the survey.

Perceptions are affected by individual-specific characteristics, such as income, gender, education, and by group-specific feature, as social norms, cultural factors, etc.
Moreover, such as for the first subgroup of surveys, reported subjective satisfaction and wellbeing may depend on the order of questions, the wording of questions, scales applied, actual mood, and the selection of information processed (Frey and Stutzer, 2002). Some of these factors could be group specific. For example, in some countries university grades are expressed o a scale from one to twenty. In a cross-country survey, respondents who attended university in those countries will have a different perception of the scale compared with individuals graduated elsewhere and with people who did not attend university at all.

Hence, surveys of this kind can provide reliable information only if, in an *ex post* analysis, are kept into account all those characteristics which are peculiar of each group of respondents and relevant in the choice process. Bertrand and Mullainathan (2001) and Grassi and Puglisi (2007) suggested econometrical techniques for isolating country-specific effect, in order to assure a reliable interpretation of surveys conducted in different countries.

4. Conclusions.

The aim of SP analysis is to provide policy makers with some information about individuals' preferences. The goodness of the policy chosen crucially depends on the reliability of SP.

We classified SP methods in two big families. The first one comprehends all those cases in which agents are required to replicate their market behaviour by choosing among different hypothetical option the preferred one. In this case the issue of reliability can be analysed from two different perspectives.

According to mainstream economics, agents' behaviour is characterised by unbounded rationality, unbounded willpower and unbounded selfishness. Then, the sole source of bias is represented by the intentional choice of some people to misreport their preferences in order to affect the policy chosen in a direction they find beneficial. Without any monetary mechanism attached to the process of revelation of preferences, dominant truth-revealing strategies do not hold. However, by adopting the model about informative lobbying (Crawford and Sobel, 1982) can be shown that in some cases the

policy-maker finds still convenient to ask agents about their preferences because, by doing so, the expected welfare loss is not lower than the one he would obtain by setting the policy chosen at random. In any case it is not possible for the researcher to detect, ex-post, misreported preferences, unless they violate assumptions about logical consistency.

The violations of the assumption which define the *homo-economics* abstraction have been identified as a relevant source of unreliability by some authors (Forster and Mourato, 2002). This approach is questionable. By observing inconsistent preferences we are admitting that some agents behave irrationally, which turns to be a rejection of our theory. This reasoning reflects the circularity of some arguments used in order to give empirical justification and universal validity to the *homo-economicus* portrait. Some authors (Wong, 1978) showed how both ordinal utility theory and the Samuelson programme failed to provide empirical verification of these assumptions. Hence, if they are not always consistent with the maximisation of individual preferences, there is no reason why individual preferences should be consistent with this set of assumptions.

The so called behavioural programme (Mullainathan and Thaler 2000) constitutes the second perspective from which we can analyse the problem of SP reliability. While Samuelson was looking for a justification of ordinal utility theory based on observations of market behaviour, the behavioural programme focuses on those market behaviours which do not respect the assumptions of mainstream economics. According to this approach two sources of bias can affect SP results. Agents may still have an incentive to falsify their choices because, as in the previous case, they want to manipulate the policy chosen in a certain profitable direction. They could also fail to correctly report their preferences, because of bounded rationality. This occurrence is now acceptable, since our theory does not assume individuals to be fully rational.

The latter consideration implies a new objection to the possibility to test the logical consistency of SP results. If we observe irrational choices in the SP exercise, how could we pretend to exclude them, if also in the real world individuals may behave in the same way? Moreover, if the assumptions about consumers' behaviour does not have universal validity, on which basis could we distinguish between consistent and inconsistent statements?

The answer concerns the peculiarities of SP exercises compared with real market situations. Fictitious market scenarios provide a complete description of the economic environment. Then, utility and happiness coincide: since the alternatives are in general fully represented by their attributes, there is no reason to make allowance for unobservables, such as moral issues or ethical considerations. In such a context the LC test can be based on the axioms of rational choice, if the consistency with those rules is relevant, without any doubt, in explaining whether people choose the options they will most enjoy.

The last section was devoted to the last family of surveys. In this case the object of the interview defines two subgroups of surveys.

In the first case individuals are asked to reveal their actual attitudes, as the number of hours spent watching TV. The peculiarity of such questions is the existence of a true and objective answer, for example the real number of hours spent watching TV. Since the objective answer cannot be observed by the researchers, the latter has to trust the respondents, who may have some incentives to misreport their behaviour or could be affected by several factors (Bertrand and Mullainathan, 2001).

The second subgroup of surveys refers to those analyses in which individuals are asked to report their feelings or their satisfaction with a certain aspect of their life. With respect to the previous case perceptions are based on personal perceptions and evaluations which depend on several unobservable and subjective factors. In the last decades the interest for this kind of studies largely increased and policy-makers are often interested in public opinion surveys since they may be interested in choosing those policies which maximise consumers' satisfaction. Then, the connection between satisfaction and objective quality is particularly relevant, especially when individuals have poor information on which base their judgements.

In this context, the issue of reliability involves not only the concept of objectivity with respect to "real" quality, but also the problem of comparability between different groups of individuals, since perceptions are affected by both individual-specific characteristics, such as education or income and by group-specific feature, such as social norms and cultural factors.

Notes

[1] Actually when Δ is equal to the boundary level the agent is indifferent between a fair statement and an unfair answer. However, we assume that he prefers to behave honestly because it is morally preferable.
[2] While behavioural economists tend to generalize this finding, some authors showed resistance to accept it as a systematic, non-negligible departure fro the traditional framework. For a critical review of the literature see Plott and Zeiler, 2004.
[3] In "A Reconsideration of the Theory of Value" (1934) Hicks and Allen talk about an increasing marginal rate of substitution, since they simply look at the change in utility the other way up the indifference curve. In the following publications they change the terminology and substitute *increasing* with *decreasing*. Here Samuelson, who refers to the 1934's article, still uses the former terminology.
[4] Wong refers to the WARP but the same situations can be adapted to the set of assumptions about the *homo economicus* profile.
[5] For a brief review see Betrand and Mullainathan (2001) and Frey and Stutzer (2001).
[6] Agents' statements can be still affected by cognitive bias, incentives to misreport their opinions, influence of cognitive factors, etc. Here we are assuming that individuals are honestly reporting their perceptions and that the latter are not affected by any source of bias.
[7] In this example we are still assuming agents to exactly report their satisfaction.

References.

Allais M., 1953, *Le Comportement de l'Homme Rationnel Devant le Risqué, Critique des Postulats et Axioms de l'Ecole Americaine*, Econometrica, 21, pp.503-546.

Bates J., 1988, *Comparison of RP and SP models of travel behaviours*, Journal of Transport Economics and Policy, Vol. XXII No. 1 January 1988.

Bentham J., 1789, *The Principles of Morals and Legislation,* edited by Burns J.H., Hurt J.L.A., Oxford University Press, 2005

Bertrand M., Mullainathan S., 2001, *Do People Mean What They Say? Implications for Subjective Survey Data*, Ameican Economic Review, 91(2), pp. 67-72.

Bonsall P., 1985, *Transfer price data – Its definition, collection and use*, in Ampt E.S., Richardson A.J., Broeg W., *New survey methods in transport*, VNU Science Press, Utrecht.

Bordignon M., Colombo L., Galmarini U., 2005, *Fiscal Federalism and Endogenous Lobbies' Formation*, CESifo Working Paper No. 1017.

Bowles S., 2004, *Microeconomics: Behaviour, Institutions and Evolution*, Princeton University Press.

Camerer C., 1995, *Individual Decision Making*, in Kagel J., Roth A.E., Handbook of Experimental Economics, Princeton University Press, pp. 587-703.

Clarke E.H., 1971, *Multipart pricing of public goods*, Public Choice, 11:17–33,

Crawford V.P., 1982, *A theory of disagreement in bargaining*, Econometrica, vol. 50, Nov. 1982, pp 607-638.

Crawford V.P., Sobel J., 1982, *Strategic information transmission, Econometrica* , vol. 50, Nov. 1982, pp 1431-1451

Ellsberg D., 1961, *Risk, Ambiguity, and the Savage Axioms*, Quarterly Journal of Economics, 75 (4): pp. 643-669.

Felli L., Merlo A., 2000, *Endogenous Lobbying*, C.V. Starr Center for Applied Economics, New York University, series Working Papers with number 00-04.

Frey B.S., Stutzer A., 2002, *What Can Economists Learn from Happiness Research?*, Journal of Economical Literature, 40(2), pp. 402-435.

Gilbert D.T., Gill M., Wilson T., 1998, *How do we know that we will like? The informational basis of affective forecasting*.

Gilbert D.T., Wilson T., 2000, *Miswanting: Some problems in the forecasting of future affective states*, in Forgas J., *Thinking and feeling: The role of affect in social cognition*, Cambridge University Press, pp.178-197.

Gollier C., 2003, *Daniel Kahnemann et l'analyse de la décision face au risque*, Revue d'Economie Politique, vol.3, pp.337-347.

Grassi S., Puglisi R., 2008, *Regulation and Consumers' Satisfaction from Public Services: an Individual Fixed Effect Approach*, Università degli Studi di Milano DEAS Working Paper June 2008.

Groves T., Loeb M., 1975, *Incentives and public inputs*, Journal of Public Economics, vol. 4 pp. 211–226.

Hensher D., 1994, *Stated Preferences Analysis of Travel Choices: The State of Practice, Transportation*, vol. 21, pp. 107-133.

Hicks J.R.,(1939), *Value and Capital*, Oxford: Clarendon Press, 2nd edition, 1946.

Hicks J.R., Allen R.G.D., 1934, *A Reconsideration of the Theory of Value*, Economica, vol.1, pp.52-76.

Hindriks J., Myles G.D., 2006, *Intermediate public economics*, MIT Press.

Houthakker H.S., 1950, *Revealed Preferences and the Utility Function*, Economica, Vol. 17, pp.159-174.

Houthakker H.S., 1961, *evealedThe Present State of Consumption Theory*, Econometrica, Vol. 29, pp.704-740.

Hsee C.K., 2000, *Attribute Evaluability and its Implications for Joint-Separate Evaluation Reversals and Beyond*, in Kahneman D., Tversky A., *Choices, Values and Frames*, Cambridge University Press, pp. 543-563.

Kahneman D., 2003, *Maps of Bounded Rationality: Psychology for Behavioral Economics, American Economic Review*, American Economic Association, vol. 93(5), pp. 1449-1475.

Knetsch J.L., Tang F.F., Thaler R.H., 2001, *The Endowment Effect and Repeated Market Trials: Is the Vickrey Auction Demand Revealing?*, Experimental Economics, 4(3).

Lindhal E., 1958, *Just taxation – a positive solution***,** in Musgrave R.A., Peacock A.T., Classics in the Theory of Public Finance, London: Macmillan.

Loewenstein G., O'Donoghue T., Rabin M., 2003, *Projection Bias in Predicting Future Utility*, The Quarterly Journal of Economics, 118:4, pp. 1209-1248.

Mas Colell A., Green J., Whinston M.D., 1995, *Microeconomic Theory*, Oxford University Press.

McFadden D., Bemmor A.C., Caro F.G., Dominitz J., Jun B., Lewbel A., Matzkin R.L., Molinari F., Schwarz N., Willis R.J., Winter J.K., 2005, *Statistical Analysis of Choice Experiments and Surveys*, Marketing Letters, 16 (3/4), pp. 183-196.

Mill J.S., 1859, *On Liberty,* edited by Himmelfarb G., Penguin Classics, 1985.

Mitchell G., 2004, *Libertarian Paternalism Is an Oxymoron*, Northwestern University Law Review, Vol. 99, No. 3.

Morewedge C. K., Gilbert D. T., Wilson T. D., 2005, *The least likely of times: How remembering the past biases forecasts of the future*, Psychological Science, 16(8),pp. 626-630.

Nisbett R.E., Kanouse D.E., 1968, *Obesity, Hunger and Supermarket Behaviour*, Proceedings of the Annual Convention of the American Psychological Association, 3, pp. 683-684.

Pareto V., 1909, *Manuel d'économie politique*, Paris: Giard, 2nd edition, 1927.

Plott C.R., Zeiler K., 2004, *The Willingness to Pay/Willingness to Accept Gap, the "Endowment Effect", Subject Misconceptions and Experimental Procedures for Eliciting Valuations*, American Economic Review.

Rabin M., Charmes G., 2000, "*Social Preferences: A Model and New Evidence*,". Department of Economics working paper, University of California, Berkeley.

Rabin M., 2001, *A Perspective in Psychology and Economics*, UC Berkley Working Paper no. E02-313.

Riley J., 1988, *Liberal Utilitarianism – Social Choice Theory and J.S.Mill's Philosophy*, Cambridge University Press.

Robinson J., 1962, *Economic Philosophy*, London: Watts.

Rothkopf M.H., 2007, *Thirteen reasons why the Vickrey-Clarke-Groves process is not practical*, Operation Research, Vol. 55, No. 2, pp. 191-197.

Ryan A., 1984, *Utilitarianism and other essays*, London: Penguin Books, pp. 11.

Saelensminde K., 2003, *The impact of choice inconsistencies in stated choice studies*, Environmental and resource economics, 23, pp. 403-420.

Samuelson P., 1938, *A Note on the Pure Theory of Consumer's Behaviour*, Economica, Vol. 5, pp. 61-71.

Samuelson P., 1948, *Consumption Theory in Terms of Revealed Preferences*, Economica, Vol. 15, pp. 243-253.

Samuelson P., 1954, *The pure theory of public expenditure*, Review of Economics and Statistics 36, pp. 387-389.

Schkade D.A., Kahneman D., 1998, *Does living in California make people happy? A focusing illusion in judgements of life satisfaction*, Psychological Science, 9, pp. 340-346.

Sen A.K., 1977, *Social Choice Theory: a Re-examination*, Econometrica, 45, pp. 53-89.

Sen A.K., 1980, *Plural utility*, Proceedings of the Aristotelian Society, 81, pp. 183-215.

Sunstein C.H., Thaler R.H., 2003, *Libertarian Paternalism*, American Economic Review, n.2, pp.175-179.

Sunstein C.H., Thaler R.H., 2003, *Libertarian Paternalism is not an Oxymoron*, The University of Chicago law Review, 70(4), pp.1159-1202.

Sunstein C.H., Thaler R.H., 2008, *Nudge – Improving Decisions About Health, Wealth and Happiness*, Yale University Press.

Thaler, R.H., 1988, *The Winner's Curse*, Journal of Economic Perspectives, Volume 2, no.1.

Tversky A., 1972, *Elimination by Aspects: A Theory of Choice*, Psychological Review, 76, pp. 31-48.

Wong S., 1978, The *Foundation of Paul Samuelson's Revealed Preference Theory*, Routledge and Kegan Paul, London.

Logical consistency in choice experiments: an application to the value of travel time.

Abstract. Stated preferences methods (SP) offer a direct approach to estimating willingness to pay (WTP) for changes in provision of non-market goods as, *e.g.,* the value of time. These methods are commonly used in Cost Benefit Analysis (CBA) and in project evaluations. The goodness of the estimation of shadow prices crucially depends on the reliability of collected answers. In other words, it is subordinated to the occurrence that people would reproduce in a real context the same behaviour expressed into the questionnaire. As pointed out by some researchers (Forster and Mourato, 1997), in many SP studies a certain number of choices seem not to respect basic assumptions of rationality. The aim of a logical consistency (LC) test consists in verifying consistency with the prediction of the economic theory. Following previous literature about the same topic, the research questions addressed by this work is dual. In the first place we want to investigate the occurrence of inconsistent choices and their effect on WTP estimations. The second issue concerns the relationships between problematic choices and respondents' individual characteristics.
Hence, a LC test has been applied on the results of a survey aimed to the evaluation of the value of travel time (VOT), expressly carried out for this research. Results show how nearly a quarter of respondents failed to provide coherent answers. The inclusion of inconsistent choices in the evaluation of travel time has a significant effect on WTP estimates. The probability to choose inconsistently seems to depend positively on age.

Contents.

1. Introduction. 46
2. Time allocation theory. 47
3. Methodology. 49
4. Survey design. 52
5. The problem of LC. 52
6. Results. 57
7.Conclusions. 60
Notes 61
Appendix. 62

1. Introduction.

The term Stated Preferences (SP) refers to a family of techniques[1] which uses individual respondents' statements about their preferences in a set of options to estimate utility functions. The options are typically descriptions of situations or contexts constructed by the researcher. In a choice experiment a scenario refers to a choice set, and the respondent is merely asked to select his preferred option from the alternatives of the set. This method corresponds with the usual discrete choice (Revealed Preferences model, RP) approach, except that both the alternatives and the responses are hypothetical. From '70s on, in many countries it is common practice to estimate the shadow price of non-market goods, such as time and pollution, trough SP analysis. These estimations are widely implemented in Cost-Benefit Analysis (CBA) appraisals and project evaluations. Despite the number of advantages relative to RP methods[2], also with SP methods researchers have to deal with some issues. Several sources of bias could affect the output of the estimation.

The structure of the questionnaire, the order of hypothetical alternatives and the attributes' level could have a significant effect on both respondents' choices.
Some authors investigated these issues and provided new design techniques for SP studies. Mc Fadden (1986) observes how factors such as learning, boredom, or anchoring to earlier tasks may distort the measurement of preferences. Fowkes et al. (2000) and Bates (1986) examined the effect of questionnaire designs on the WTP estimation and discussed the goodness of some characteristics, such as orthogonality[3], in SP surveys.

Another source of bias derives from the fact that respondents may deliberately give biased answers, in the hope of affecting the outcome of the analysis (policy-response bias) or of easting their existing behaviour in a better light (justification bias)[4]. These kinds of bias are not easily identifiable, since we do not observe "real" preferences and consequently, we have to trust in respondents' statements.

Even with a well suited questionnaire, and under the assumption that all respondents will seriously and honestly face the SP exercise, a source of bias is still represented by the cognitive difficulty associated with multiple complex choices between scenarios with many attributes and levels. The relevance of this noisy element varies among individuals, since some of them can find the task easier than others, due to a combination of observable (age, education, profession) and unobservable factors.

Summing up all this factors, SP studies can be biased by an improper design of the questionnaire, by a strategic behaviour of the respondents and by the lower ability of some individuals to answer the questionnaire. Unfortunately, most of these factors cannot be detected and we are not able to check whether collected SP coincide with RP.

However, since economic theory provides us some suggestions about individuals' behaviour, we can check the consistency of these assumptions with SP. As Forster and Mourato (1997, 2002) observe, the existing literature contains few systematic attempts to examine whether the collected answers satisfy certain condition of logical consistency (LC), in accordance to the predictions of economic theory. Their contribution investigates the occurrence of inconsistent choices in a contingent ranking experiment. Saelensminde (2002) replicates the exercise in a choice experiment aimed at the estimation of the value of travel time (VOT).

Following Saelensminde (2002), the purpose of this paper is to observe the effect of the inclusion of inconsistent choices on VOT evaluations and to underline the connections between the tendency to choose inconsistently and individual characteristics. Compared with previous literature the LC test is less restrictive, as it will be discussed in the next sections. The study uses data from a survey specifically conducted for this work and aimed at the evaluation of the VOT for train travellers between Turin and Milan. A sample of 407 travellers has been asked to answer a questionnaire composed by a set of nine couples of travel alternatives, each of them including only two attributes, in order to simplify the task for the respondents. The last section looks for a relationship between the tendency of choosing inconsistently and some observable socio-economic variables.

2. Time allocation theory.

The first attempt to consider time as a commodity goes up to Becker (1965). He suggested that individual utility does not derive from goods consumed directly but from "final goods" Z_i,

$$Z_i = f_i(x_i, t_i)$$

defined as a function of a vector of market goods and a vector of time inputs used in producing the *i*th commodity. Then, time enters the utility function through Z_i. In this model the time constraint is not explicitly defined. It is incorporated into the traditional budget constraint as time can be converted into goods by using less time at consumption and more at work. Hence, the consumption activity has a time cost, equal to the cost of not earning money. Then, in its first acceptation, the VOT is equal to the wage rate, independently from the allocation of time between different kinds of activities, pleasant or not.

Johnson (1966) modifies the structure of the utility function which depends, in his model, on three separate attributes: money income, time spent at work and hours of leisure. In this way he shows that the value of time plus the subjective value of work, defined as the ratio between the marginal utility of work to the marginal utility of income. As the former is assumed to be negative while the latter is positive, the value of leisure or any use of time will be less than the money wage rate. In other words, the

intrinsic value of saving time is equal to the difference between the wage rate and the disutility of labour.

Oort (1969) includes travel time in the utility function as well. The VOT is then equal to the sum of the intrinsic value of saving time, which is the definition given by Johnson (1966), and the marginal utility of the transport activity itself, assumed to be negative.
In both Johnson (1966) and Oort (1969) model, beside a budget constraint, a time constraint shows that the sum of leisure and work time cannot overcome the daily time endowment.

The first attempt to generalize these results is due to DeSerpa (1971). In this model he considers a set of commodity bundles
$x = (x_1,...,x_n,t_1,...,t_n)$
Where x_i denotes the quantity of *i*th consumption good and t_i denotes the time allocated to the ith good. One of this activity is work and it is denoted by t_w, which enters separately in the direct utility function: $U(x,t,t_w)$. Defining w as the wage rate and y as the amount of income available from non-work sources, the traditional budget constraint results

$$px = w \cdot t_w + y \qquad (1)$$

According to the previous literature, time is constrained as well and following relationship between total available time (T) and time spent in various alternatives must hold

$$T \geq \sum t_i + t_w \qquad (2)$$

At this stage DeSerpa notices that some activities, e.g. a trip, require a minimum amount of time (t_i^*) for their consumption. In order to make more realistic the model he introduces a number of time constraints equal to the number of activities

$$t_i > t_i^* \qquad (3)$$

$$t_w > t_w^* \qquad (4)$$

Hence, the utility function must be maximized subject to (1), (2), (3) and (4).
Writing the Lagrangean as

$$L = U(x,t,t_w) + \lambda(wt_w + y - px) + \mu(T - \sum t_i - t_w) + \phi(t_w - t_w^*) + \sum \psi_i (t_i - t_i^*)$$

FOCs follow:

$$\begin{cases} \partial U / \partial x_i - \lambda p_i = 0 \\ \partial U / \partial t_j - \mu + \psi_j = 0 \\ \partial U / \partial t_w + \lambda w - \mu + \phi = 0 \end{cases}$$

The marginal valuation of time spent in activity *j* is the ratio of the marginal utility of activity *j* to the marginal utility of income:

$$\frac{1}{\lambda}\frac{\partial U}{\partial t_j} = w + \frac{1}{\lambda}\frac{\partial U}{\partial t_w} + \phi / \lambda - \psi_i / \lambda$$

where the first term on the right hand side is the VOT as defined by Becker (1965), while the first two terms represent the result of Johnson (1966).

The same formula can be obtained by dividing the second FOC by λ:

$$\frac{1}{\lambda}\frac{\partial U}{\partial t_j} = \mu / \lambda - \psi_i / \lambda$$

From this formulation DeSerpa defined three types of values and the relationship established among them. The *VOT as a resource* (μ / λ) is defined as the value of extending the time period when the time constraint is not binding. It represents the WTP for an increase in the total time budget. The second is the VOT allocated in a certain activity (*VOT as a commodity*), given by the rate of substitution between that activity and money in the utility function. This would be equal to (μ / λ) only if the individual assigns more time to an activity than the minimum required. If it is the case, this particular activity will be classified among leisure activities for which, as a consequence, the value of saving time is zero. The third concept is the value of saving time in activity *i* (*value of transferring time*), defined as the difference between the marginal valuations of time spent in the activity *i* and the resource value, equal to ψ_i / λ. It is this concept which is conventionally referred to as the VOT in transport appraisals (Layard and Glaister, 2003).

3. Methodology.

3.1. Discrete travel choice.

Discrete choice models are the most common type of travel demand models. Empirical measurement, such as RP methods, is confined to situations involving choices between two or more mutually exclusive alternatives. The (indirect) utility level attached to each alternative is usually represented a s a linear combination of cost and characteristics of each alternative and other general effects, e.g. depending on socio-economic variables for each group of individuals. But the revealed choice cannot be entirely explained by observable factors, since other unobserved variables, as personal tastes, could enter the consumer's decision process. In other words, there is a random element which cannot be determined (Manski, 1977). Thus the basic random utility formulation is that:

$$U_i = U\left(V_i, e_i\right)$$

which is commonly simplified to the additive formulation

$$U_i = V_i + e_i$$

Where V_i is the deterministic part of the utility function and, as a further simplification, is usually specified as a linear-in parameters function of a coefficient vector β and a vector of explanatory variables. Whether the model is calibrate on individual observation, econometrical technique estimates that set of coefficients which, when

inserted into the formula of V , maximises the joint probability across all the observations of the choices made by the respondents. This allows values of travel time to be estimated as the ratio of the estimated utility weights of travel time to that of cost.
For example, consider the case of n individuals who have to choose between i alternatives (i = car, train) defined only by two attributes: travel time and cost. Then, the observable part of the indirect utility V of the alternative chosen i can be written as

$$V_{ni} = \beta_0 + \beta_{time} time_i + \beta_{\cos t} \cos t_i$$

This specification is largely used in transportation appraisals in order to evaluate the VOT as the ratio between the coefficients related to time and cost. As Domencich and McFadden (1975) show, based on the assumption that the set of e_i are independent from irrelevant alternatives (IIA) with a Gumbel distribution, leads to the multinomial logit model. Hence, the probability of the car to be chosen (car is set equal to 1 and train is equal to 0) can be written as:

$$\Pr(y_n = 1) = \frac{e^{V_{ncar}}}{e^{V_{ncar}} + e^{V_{ntrain}}} = \frac{1}{1 + e^{-(V_{ncar} - V_{ntrain})}}$$

where y_n is the choice stated by the n*th* respondent
In this work we applied to our analysis a multinomial logit model using package NLOGIT developed by Limdep.

3.2.LC and individual characteristics.

The relationships between inconsistent choices and individual characteristics has been analysed from two different perspectives.

In the first one, a binomial logit model is applied. Logical skills act as the latent, unobserved variable in the model

$$y^* = X\beta + e$$

where e is a continuously distributed variable independent of X and the distribution of e is symmetric about zero. Instead of the latent variable y^* we observe a binary variable y_i, which is equal to 1 if the i*th* respondent chose inconsistently and is equal to 0 otherwise. If the cumulative distribution function of e is assumed to follow a standard logistic distribution, such as

$$G(e) = \Lambda(e) \equiv \exp(e)/[1 + \exp(e)]$$

the traditional binomial logit model follows.

Another way to look at the same issue involves Bayesian networks (Pearl, 2000). These methods are largely used in biology and medicine literature, whilst a few applications have been carried out in economic literature.

A Bayesian Network (BN) is a graphical model for probabilistic relationships among a set of variables. Each variable is represented by a node in a graph. The direct dependencies between the variables are represented by directed edges between the corresponding nodes and the conditional probabilities for each variable (that is the probabilities conditioned on the various possible combinations of values for the immediate predecessors in the network) are stored in potentials (or tables) attached to the dependent nodes. Information about the observed value of a variable is propagated through the network to update the probability distributions over other variables that are not observed directly. Using Bayes' rule, these influences may also be identified in a 'backwards' direction, from dependent variables to their predecessors (Heckerman, 1996).

Hence, a BN consists of the following (Jensen and Nielsen, 2007):

1. a set of variables and a set of directed edges between variables;
2. each variable has a finite set of mutually exclusive state;
3. the variable together with the direct edges form an directed acyclic graph (DAG); a directed graph is acyclic if there is no directed path $A_1 \to ... \to A_n$ so that $A_1 = A_n$;
4. to each variable A with parents $B_1,...,B_n$ a conditional probability table $P(A|B_1,...,B_n))$ is attached;
5. The joint probability $P(A_1,...,A_n)$ factorises along the graph: that is if $A_\Pi(i)$ denotes the set of parents of the variable A_i (and the empty set if A_i does not have any parents), then

$$P(A_1,...,A_n) = \prod_{i=1}^{n} P(A_i | A_\Pi(i))$$

In empirical works, researchers handle a sample of cases from a network N over the universe U. Starting from this sample, the object consists in reconstructing the BN from the cases. In order to reach this goal, two methodologies are usually applied.

The constraint-based methods establish a set of conditional independence statements holding for the data, and use this set to build a network corresponding to the conditional independence properties determined. In other words, as a first step the skeleton of the BN is determined by analysing conditional independences among variables and then the arcs are directed.

The score-based methods produce a series of candidate Bayesian networks, calculate a score for each candidate, and return the candidate of highest score. This score represents an indicator for the probability that our sample could have been generated by each BN candidate. The network which fits better the sample will receive the highest score and will be chosen. Naturally, we are assuming that our dataset constitutes a correct representation of the "true" network, N.

Several software employ algorithms and models devoted to BN's estimations. In this application we use the package *R* and the library *bnlearn* (Scutari, 2010). Compared with other software, the advantage of using this package relies in the possibility to perform both constrained-based and score-based methods on our sample, testing several and alternative algorithms. The main disadvantage is connected to the impossibility to mix in our data set continuous variables (such as age) and factors. Even if some libraries (*deal*, Bottcher and Dethlefsen, 2007) handle networks with mixed variables, their learning procedure is still experimental and hardly applicable to complex models.
For this reason all variables have been converted into factors, as it will be explained with more details in the sixth paragraph.

4. Survey design.

The data set used in this paper includes the results of a survey about the VOT, specifically performed for this research in June and July 2008.
A sample of 407 second-class travellers[5] on the regional train between Turin and Milan has been asked to answer a questionnaire, whose design is shaped based on other studies conducted in a similar environment (TRT, 2001).

Respondents had to choose between two travel alternatives in nine different hypothetical scenarios. Each journey is described by two attributes: ticket price and travel time. The transport mode and its characteristics (comfort, cleanness, etc.) are assumed to be constant across alternatives. Respondents have also been asked about some individual characteristics such as age, gender, income, level of education and use of the transportation mode (journeys per month), reason of the journey (work or leisure trips). A more detailed description of the questionnaire is available in the appendix, jointly with some descriptive statistics about the respondents.

At the end of the questionnaire, after the SP exercise, some questions tried to investigate travellers' behaviour and their attitude to attach a monetary value to the time saved. More details about this section are given in the next paragraph.

5. The problem of LC.

When performing a SP analysis, the final goal of the researcher is to get a perfect substitute for unobserved RP. As a consequence, the reliability of collected data depends on their capability to predict real preferences. Since the latter are unknown, it is not possible to compare statements with actual behaviours. However we can, *ex ante*, make some assumptions about individuals' market behaviour and, *ex post*, check whether these hypothesis hold across statements. If our results are consistent with our assumptions we can suppose individuals properly reported they preferences. If it is not the case, we will doubt about the reliability of SP and we could decide to drop from our sample those choices which are inconsistent with our hypothesis.

The possibility to check about LC crucially depends on the design of the questionnaire. Foster and Mourato (2002) look at three different aspects of LC within the context of a contingent ranking experiment: dominance, rank consistency and transitivity[6]. Summing up all inconsistencies, Forster and Mourato (2002) detected problematic choices in 41 cases out of 100.

Compared with choice experiments, a limitation of the ranking approach lies in the added cognitive difficulty associated with many attributes and levels. Consequently, it is easier to find irrational choices among the answers given by the respondents. On the other hand, the construction of the questionnaire allows to check for several kinds of inconsistent choices, while in choice experiments only dominance problems can be observed, since they involve just one or two hypothetical comparisons.

Before testing the reliability of SP results we have to check the robustness of our assumptions. In other words, keeping in mind our very final objective, we must define our hypothesis consistently with individuals' actual behaviour.

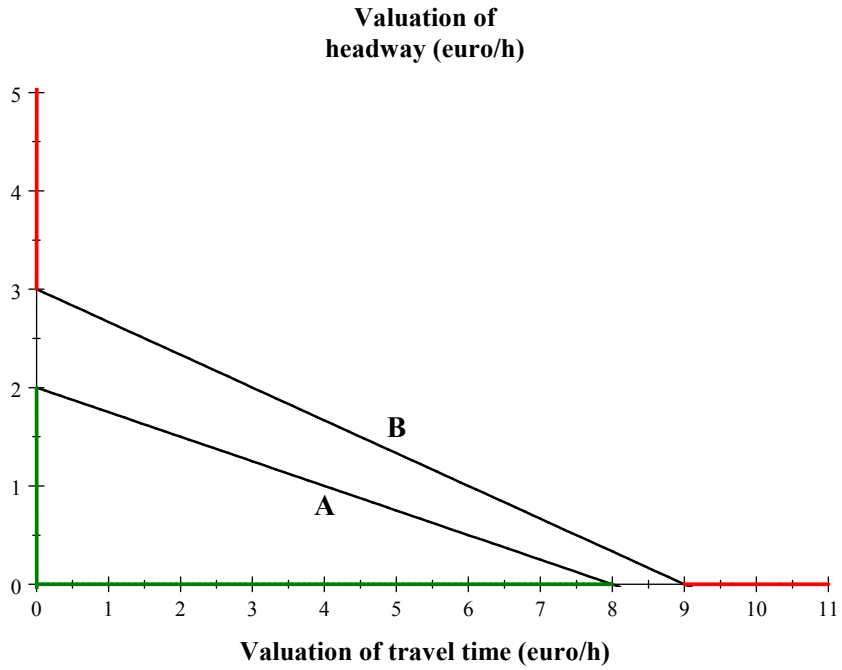
Saelensminde (2002) tries to investigate the impact of choice inconsistencies by using the results from a VOT study conducted in Denmark. In this experiment travel alternatives are described by three attributes: price, travel time and headway, defined as the number of hours between each departure. Inconsistencies are identified by considering the linear combinations of the valuation of travel time and headway and, according to the continuity axiom, it is assumed that the respondents trade-off gains in travel time against losses in headway, and vice versa. Empirical results show how, given this formulation of the LC test, about 75.5 per cent of train travellers chose inconsistently.

For example, suppose to present travellers with two couples of alternatives. In each case (A and B), travellers are asked to choose between two scenarios the one they prefer. Following Saelensminde (2002), if the respondent in choice A chooses the cheapest option (scenario 1), then he/she is not willing to pay 2 euro more for a reduction in travel time of 15 minutes and a decrease in headway of one hour.
This behaviour indicates that his/her valuation of reduced headway is less than 2 euro per hour and the valuation of reduced travel time is below 8 euros per hour, as depicted by the green line in figure 1.

Following the same reasoning, if the respondent chooses the most expansive option in choice B (scenario 2), his/her WTP for time will be higher than 9 euro/hour and his/her WTP for a reduction in headway will be above 3 euro/hour, as shown by the red lines in figure 1. The simultaneous choice of the first scenario in (A) and of the second one in (B) implies a failure of the consistency test, since preferences are contradictory. Two observations can be moved to this definition of LC.

Figure 1. Logical consistency: choice A *vs* choice B.

	Choice A		Choice B	
	Scenario 1	*Scenario 2*	*Scenario 1*	*Scenario 2*
Travel time	1h 45	1h 30	1h 45	1h 25
Price	8 euro	10 euro	8 euro	11 euro
Headway	6 hours	5 hours	6 hours	5 hours



First of all behavioural economic literature pointed out how individuals often base their choices on rules of thumb, as a way to economize on cognitive faculties (Mullainathan and Thaler, 2000). In the case represented in figure 1 we are assuming agents' WTP to be constant and unrelated to the size of the time saving offered (respectively 15 and 20 minutes in choice A and B). In our questionnaire we tried to investigate the realism of this assumption. Train travellers between Turin and Milan (length of journey: 1h 45') have been asked through three direct questions about their WTP.

The purpose of the first question was just to let them understand the object of the interview: *if I offered you another ticket, a little bit more expensive compared to the one you bought today in order to save a minute of travel time, would you buy it?* Obviously all respondents answered "no".

The second question was the following: *does it exist a minimum amount of time that I should offer you, in order to convince you to buy a more expensive ticket with respect to the one you bought today?* This time answers differ meaningfully among respondents (table 1). None of the respondents would accept an increase in the price of the ticket

without a correspondent reduction in travel time of, at least, ten minutes. 16 per cent of the sample is not able to quantify the minimum amount of travel time that would justify a higher price. Then, referring to figure 1, in a RP context we could observe some individuals who accept the expensive option in choice B but reject the cheapest one in choice A. As a consequence such a choice cannot be defined as inconsistent.

Moreover, in the model proposed by DeSerpa (1971) each activity requires a minimum amount of time for its consumption. In a choice experiment respondents are asked about their WTP for a decrease in travel time. In accepting or refusing a certain option, travellers evaluate the opportunity to spend the time saved into another activity, which is constrained by a minimum amount of time. If the time saving is too small they will not find anything else to do, and the utility gained will simply be equal to the avoided disutility of travel. For example, no one would pay for a one minute decrease in travel time.

If the size of the time saving increases, agents will be able to invest this time into other consumption or work activities, so that the gain in utility will be equal to the avoided disutility of travel plus the utility connected with the new occupation. In other words, we are assuming that the value of transferring time is not constant, since time has no value *per se* but only in relation to its use. Again, the choice discussed in figure 1 cannot be defined as inconsistent.

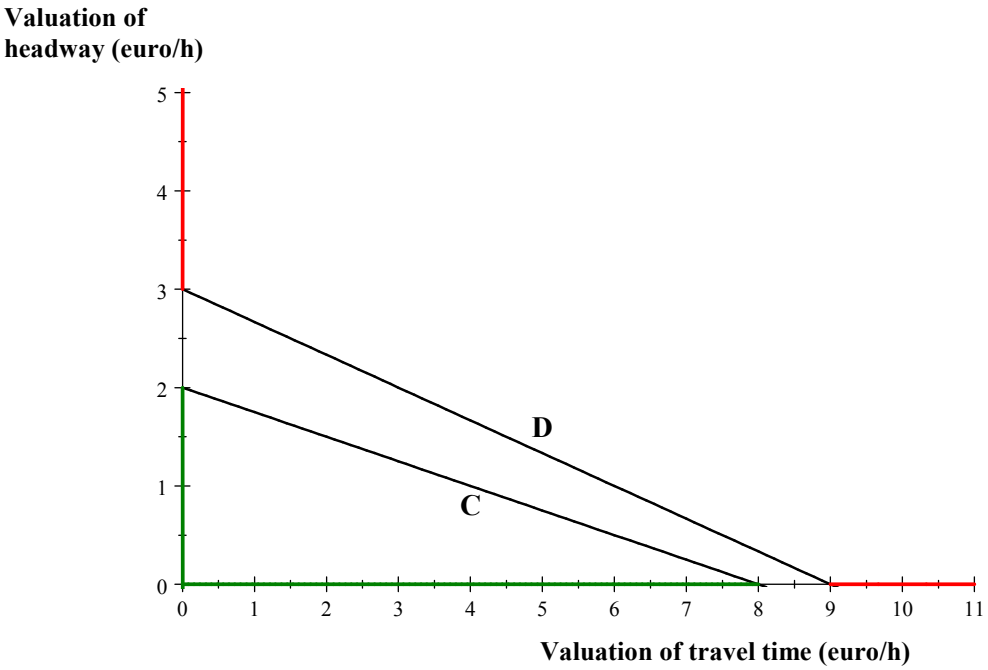
The same reasoning holds for the size of monetary costs. Consider for example choices C and D. Figure 2 shows how WTP changes if the respondent accepts or refuses each option. According to Saelensminde (2002) this time LC would be violated by the choice of the cheapest option in C (scenario 1) and the most expansive one in D (scenario 2).

A third question tried to investigate travellers' ability to attach a monetary value to the minimum amount of time for which they would be willing to suffer an increase in the price. Question wording was: *how much, if it exists, is the maximum amount of money that you would be willing to pay in exchange for a time saving?* Answers significantly differ among individuals. This time the number of respondents is lower (340) since those who answered "no idea" in the previous question have not been asked. Some travellers are not able to quantify the threshold (15.45 per cent), some others (11.18 per cent) answer in proportion to the current price of the train ticket (table 2).

Empirical evidence seems to show the existence of thresholds, both for time and cost. Below a certain level of time and above a certain level of price travellers' WTP tends to zero. Because of this evidence, in this study LC is defined in a less restrictive way. First of all travel alternatives are characterised by two attributes instead of three as in the previous examples: travel time and price. This design significantly reduces the mental effort for the respondents. LC is not checked though WTP but by comparing attributes' levels between choices.

Figure 2. Logical consistency: choice C *vs* choice D.

	Choice C		Choice D	
	Scenario 1	Scenario 2	Scenario 1	Scenario 2
Travel time	1h 45	1h 20	1h 45	1h 30
Price	8 euro	14 euro	8 euro	12 euro
Headway	6 hours	4 hours	6 hours	5 hours



Consider for example the two following situations, A and B.

	Choice A		Choice B	
	Scenario A1	Scenario A2	Scenario B1	Scenario B2
Travel time	1h 45	1h 30	1h 45	1h 20
Price	8 euro	10 euro	8 euro	13 euro

If a respondent chooses alternative *A1*, his/her WTP is lower than 8 euro/hour. But in the second scenario he could prefer alternative *B2*, even if the value of each unit of time saved is higher (higher than 12 euro/hour), as the amount of time saved (25 minutes) could justify this behaviour. In other words, since $A1 = B1$, we could write:

$A1 \succ A2$

$B2 \succ A1$

As we can not infer that $A2 \succ B2$, the choice does not violate LC.

Consider now choices C and D.

	Choice C		Choice D	
	Scenario C1	Scenario C2	Scenario D1	Scenario D2
Travel time	1h 45	1h 30	1h 45	1h 30
Price	8 euro	10 euro	8 euro	10.50 euro

In this case the simultaneous choice of alternative C1 and D2 is not logically consistent, because in the scenario D2 the time saved is the same as in C2 (15 minutes) but the fare is higher. Since $C1 = D1$ the choice can be written as

$C1 \succ C2$

$D2 \succ C1$

But this time we know, assuming a strongly monotone preference relation, that $C2 \succ D2$. Since this relation contradicts the SP of the respondent, his/her choices violate the LC test.

The same reasoning holds if we keep constant the amount of time saved and let change the cost of the ticket. The inclusion of some dominant alternatives in the questionnaire allows checking for LC without increasing too much the difficulty level of the task. In most of cases dominant pairs are excluded from questionnaires on the ground that they do not provide any additional information about preferences.

6. Results.

The occurrence of inconsistency is verified for 105 respondents (table 3), while one third of the choices reflects lexicographic preferences and cannot be checked for LC. Lexicographic preferences characterise those individuals who always chose the cheapest option or the most expansive one in all nine couples of alternatives. The share of travellers who passed the LC test is much higher compared with the result (24.5 per cent) for train commuters in Saeleminide (2002).

Results for the estimation of VOT are reported in table 4. Here, as common practice in transportation appraisals, the whole sample (407 respondents) enters the model of specification of the utility function, without considering the positive or negative result of the LC test. Two models have been estimated. In the first one the observable part of the utility is assumed to be linear with respect of travel time and cost, as in the example in section 3. In the second model we added some individual characteristics in the estimated utility function.

Table 4 shows how the VOT derived by considering the whole sample is about 13.08 euro per hour, a value slightly lower with respect to other studies in our country (Cherchi 2003) and to the values suggested for Italy by UNITE (2001). The inclusion of individual characteristics does not change WTP estimations. Travelling for work reasons has a negative influence on utility, while concerning wealth we observe that the higher the income the higher the disutility of travel. The first column of table 6 shows

how WTP differs between different categories of travellers: as expected VOT is higher for those who travel for work reasons and for respondents who declared a net income greater than 20.000 euro per year.

The same procedure has been applied on a restricted sample, comprehensive of those respondents who passed the LC test. Results are summarised in table 5. Compared with the results summarised in table 4, the VOT for the restricted sample is lower and equal to 12.34 euro per hour. This result is consistent with the evidence found by Forster and Mourato (2003), Saeleminide (2002) and holds also with other data sets (TRT, 2001). This overall effect is due to the increase of the relative weight of lexicographic choices in our sample. As we said, about 30.7 per cent of respondents always chose the cheapest (and slowest) option or the most expensive (and fastest) one. A strong majority (85 individuals out of 125) are in the first group, i.e. they preferred the most affordable scenario in all nine choice situations. Then, by refusing the second option in choice 2 (table 8), which is the most convenient one offered in the questionnaire, they declared a WTP for travel time lower than 8.6 euro/hour. Hence, if we drop from the sample all the observations which violate the LC test, the proportion of individuals characterised by the lowest VOT increases.

This effect is shown in table 6, where the VOT for some specific groups of individuals is summarised. The exclusion of inconsistent SP has a stronger effect (about -10 per cent) on the WTP estimations for non-work trips and for travellers who declared a net income lower than 20.00 euro. This is due to the fact that in these categories a large share of respondents always chose the cheapest travel option.

Table 6 shows how the variation in VOT estimates between the full sample and the restricted one is significant. Such a difference provides important implications for CBA appraisals. In this particular case, the exclusion of inconsistent choices leads to lower estimates of benefits rising from reductions in travel time. It is not clear whether the exclusion of inconsistent choices in the model estimation is justified or not.

On the one hand, the inclusion allows[7] for violation of preference relations. This can leads to distorted evaluations of time, since some of the respondents were not able to express their preferences through the questionnaire. Hence, SP are not good substitutes for RP methods[8].

On the other hand, the exclusion of irrational choices could lead to self selection biases[9]. If the tendency to answer illogically depends on some observable factors, such as age, education or income, the use of a restricted sample could imply that the preferences of a certain category of respondent are not taken into account in the WTP estimation.

Table 7 shows the results of our attempt to investigate the relationship between inconsistent choices and individual characteristics. Two logit models have been estimated, the first one (A) includes the full sample, whilst in the second (B)

lexicographic choices have been excluded. The strongest result is the one concerning the age of respondents: younger people are less likely to choose inconsistently. Considering model A, a shift from the minimum value of age to the maximum one implies an increase in the probability of failing the LC test of about 35 per cent.
Respondents who are travelling for work reasons tend to choose more inconsistently in model A. This evidence could be explained by the assumption that people who travel for non-work reason are more relaxed and then they can pay more attention to the questionnaire. However this result is not confirmed in model B.
Respondents who chose the train because they find it somehow "better" (cheaper, faster, safer) compared with other travel options are more likely to pass the LC test.
Finally, a relation between travellers' satisfaction and probability of test failure has been found: the higher the satisfaction with punctuality the lower the probability of choosing inconsistently. Again, as an explanation we could assume that unsatisfied respondents are not really interested in answering the questionnaire, since they think that hypothetical reductions in travel time would be defeated by the inefficiency of the service.

Results showed in table 7 do not confirm the evidence from Saeleminide (2002), where the only statistically significant variable was the level of education[10].

Figure 3 reports the BN resulting from the estimation described in the methodological chapter. As we said, the package applied in this analysis does not allow mixed datasets. Hence, the unique continuous variable from our survey (age of respondents) has been recoded into four categories: under 35, 36-48 years, 49-60 years and over 60.

The network combines background information with a learning procedure generated using the Hill Climbing algorithm, a score based method implemented in the library *bnlearn*. With the expression *background information* we simply mean that at the beginning of the learning procedure we expressly made some trivial assumptions about the networks' structure. For instance, gender and age of the respondents cannot be induced by any other variable. As cited above, the library *bnlearn* implements several learning algorithms and we tested our sample on six different algorithms which apply both constraint-based and score-based methods.

Results from the logit estimations are confirmed, but the analysis of the BN provides some extra evidence about the connections among variables. Logical consistency (LOGCONS) is directly influenced by the age of the respondents and the reason of the trip (WORK), which allows to distinguish between work trips and leisure journeys. Moreover, the age affects logical consistency both directly and indirectly, through the reason of the trip.

As we said, each node has a conditional probability table attached, which describes the connection between the children node (here, logical consistency) and its fathers (age and reason of the trip). The parameters of node LOGCONS are reported in table 8. The

probability to answer inconsistently conditioned on age is higher for elderly who travel for non-work reasons compared with other age categories and with those who declared to travel for work. However, we have to keep in mind that the reason of the trip is indirectly influenced (through the level of education, EDU) by the age of the respondents.

7.Conclusions.

In this paper a test for LC has been developed on a survey of train travellers.

In general, the issue of LC in SP analysis is still largely unknown especially in choice experiment studies where, very often, the design of the surveys does not allow for consistency tests. This is mainly due to the high cost of this kind of surveys, which increases with respect of the number of questions and the dimension of the sample.

With respect to previous literature, the definition of inconsistency was much less restrictive. Moreover, the choice situations proposed in the SP exercise are characterised by only two attributes (travel time and cost), in order to reduce the complexity of the task.

Results show that, even using less constraining assumptions, a significant number of respondents (25.80 per cent) did not choose according to our LC test. The inclusion of illogical answers in the WTP valuation leads to an overestimation of the VOT.

It is doubtful whether this exclusion is correct or not, since it could imply an under-representation in the sample of certain groups of individuals which find more difficult to reveal their preferences through a questionnaire. In our case, the tendency to choose inconsistently seems to be positively related with the age of respondents.

This result is confirmed by an analysis conducted through Bayesian Networks. This methodology shows how the age of respondents has both a direct and an indirect effect on logical consistency. Elderly find more difficult to correctly report their preferences and the same holds for those who are travelling for non-work reasons. The indirect effect of age acts because individuals in the highest age group are more likely to work for non-trip reasons.

Finally, an issue which merits further research is the presence of lexicographic choices in SP studies. As we have seen, a large share of respondents (30.7 per cent) always chose the cheapest or the most expensive travel alternative. Such behaviour, which is recurrent in this kind of surveys (Saeleminide, 2002), could be explained by very low (or very high) WTP but also by assuming respondents' strategic behaviour. For example, if they are worried about an increase in train fares they could have an incentive to underestimate their VOT. Considering our sample, no evidence has been found about a relationship between this kind of choices and individual characteristics.

Notes

[1] SP techniques can be split in four categories:
choice experiments: respondents have to choose between two or more alternatives;
contingent ranking: respondents are asked to rank a series of alternatives;
contingent rating: respondents are asked to score alternative scenarios;
paired comparisons: respondents have to score pairs of scenarios on similar scale.
It is doubtful that two last methods are able to provide welfare consistent estimates (OECD, 2006).

[2] As pointed out by Kroes e Shelron (1988), RP studies are often characterised by strong correlations between explanatory variables of interesting (particularly travel time and cost). These make it difficult to estimate model parameters reflecting the proper trade-offs ratios. Moreover, RP preference methods cannot be used in a direct way to evaluate demand under conditions which do not yet exist. Finally, they require that the explanatory variables can be expressed in "objective" or "engineering" units; therefore they are normally restricted to primary service variables (such as journey time and cost) and can in practice rarely be used to evaluate the impact of changes in secondary travel variables (such as seat design and station facilities).

[3] See Bouffioux (2002) for a survey on choice experiments design techniques.

[4] These and other forms of bias response have been classified by Bonsall (1985).

[5] Only travellers who bought a full price ticket (8 euro) have been interviewed.

[6] Dominance: one alternative is said to dominate a second when it is at least as good as the second.
Rank consistency: if a respondent is asked to rank options A,B,C,D in the first question and options A,B,E,F in the second question, rank consistency requires that a respondent who prefers A over B in the first question continues to do so in the second question.
Transitivity: if a respondent expressed a preference for A over B in the first question and for B over C elsewhere, he should not express a preference for option C over A in any other question.

[7] At least with the definition of dominance used in this study.

[8] If we assume that people would always choose consistently in a real context, which is not always the case.

[9] A self selection biased is already present in SP studies since some people accept the interview and some others refuse. No evidence exists, at least in economic literature, about the relationship between the tendency to accept the interview and social-economics parameters.

[10] In Saeleminide (2002) the survey used for the regression of inconsistent choices is not the same as in the VOT experiment.

Appendix.

Table 1. **Does it exist a minimum amount of time that I should offer you, in order to convince you to buy a more expensive ticket with respect to the one you bought today?**

	Frequency	Percentage	Cumul. Perc.
10 minutes	7	1.64	
15 min.	27	6.56	1.64
20 min.	100	24.59	8.20
25 min.	33	8.20	32.79
30 min.	135	33.17	40.99
45 min.	25	6.56	74.16
50 min.	5	1.23	80.72
60 min.	7	1.64	81.95
No idea	67	16.39	83.59
Total	**407**	**100,00**	**100,00**

Table 2. If it exists, how much is the maximum amount of money that you would be willing to pay in exchange for a time saving?

	Frequency	Percentage
2 euro	13	3.86
2.5 euro	2	0.59
3 euro	39	11.59
4 euro	33	9.65
5 euro	46	13.52
6 euro	26	7.72
7 euro	14	4.12
8 euro	33	9.65
9 euro	21	6.18
10 euro	13	3.86
12 euro	2	0.59
15 euro	7	1.93
No idea	53	15.45
Answer in % of the current ticket price	38	11.18
Total	**340**	**100,00**

Table 3. Number of respondents who chose inconsistently.

	Frequency	Percentage
Inconsistent choices	105	25.80
Lexicographic choices	125	30.71
Consistent and not lexicographic choices	177	43.49
Total	407	100.00

Table 4. Valuation of travel time on the whole sample (407 respondents).

```
+-----------------------------------------------+
| Discrete choice (multinomial logit) model     |
| Maximum Likelihood Estimates                  |
| Dependent variable               Choice       |
| Weighting variable                 None       |
| Number of observations             3663       |
| Iterations completed                  5       |
| Log likelihood function       -2231.005       |
| R2=1-LogL/LogL*  Log-L fncn   R-sqrd  RsqAdj  |
| No coefficients   -2538.998    0.121   0.120  |
| Constants only    -2394.229    0.068   0.067  |
| Chi-squared[ 2]          =    326.448         |
| Prob [ chi squared > value ] =   0.000        |
| Response data are given as ind. choice.       |
| Number of obs.=  3663, skipped   0 bad obs.   |
+-----------------------------------------------+
+---------+--------------+----------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+--------------+----------------+--------+---------+
 K              0.341          0.057          5.973    0.000
 BTIME         -0.077          0.005        -16.320    0.000
 BCOST         -0.353          0.021        -16.725    0.000
```

VOT= (BTIME/ BCOST)*60 = 13.08 euro/hour

```
+-----------------------------------------------+
| Discrete choice (multinomial logit) model     |
| Maximum Likelihood Estimates                  |
| Dependent variable               Choice       |
| Weighting variable                 None       |
| Number of observations             3663       |
| Iterations completed                  5       |
| Log likelihood function       -2132.249       |
| R2=1-LogL/LogL*  Log-L fncn  R-sqrd  RsqAdj   |
| No coefficients  -2538.998  0.160  0.158      |
| Constants only   -2394.229  0.109  0.107      |
| Chi-squared[ 8]          =    523.960         |
| Prob [ chi squared > value ] =   0.000        |
| Response data are given as ind. choice.       |
| Number of obs.=  3663, skipped   0 bad obs.   |
+-----------------------------------------------+
+---------+--------------+----------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+--------------+----------------+--------+---------+
 K              1.410          0.290          4.861    0.000
 BTIME         -0.082          0.005        -16.677    0.000
 BCOST         -0.374          0.022        -17.092    0.000
 BINCOME       -0.203          0.020        -10.210    0.000
 BAGE           0.008          0.003          2.402    0.016
 BWORK         -0.232          0.085         -2.742    0.006
 BEDU          -0.038          0.069         -0.544    0.586
 BFEMALE       -0.078          0.075         -1.040    0.298
 BFREQ         -0.116          0.045         -2.590    0.010
```

VOT= (BTIME/ BCOST)*60 = 13.09 euro/hour

Table 5. Valuation of travel time on the "clean" sample (311 respondents).

```
+--------------------------------------------+
| Discrete choice (multinomial logit) model  |
| Maximum Likelihood Estimates               |
| Dependent variable              Choice     |
| Weighting variable                None     |
| Number of observations            2727     |
| Iterations completed                 5     |
| Log likelihood function       -1631.251    |
| R2=1-LogL/LogL*  Log-L fncn  R-sqrd  RsqAdj |
| No coefficients  -1890.212    0.137   0.136 |
| Constants only   -1803.763    0.096   0.095 |
| Chi-squared[ 2]           =    345.024     |
| Prob [ chi squared > value ] =   .000      |
| Response data are given as ind. choice.    |
| Number of obs.=  2727, skipped   0 bad obs. |
+--------------------------------------------+
+---------+--------------+----------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+--------------+----------------+--------+---------+
 K              0.139          0.065         2.127   0.033
 BTIME         -0.089          0.006       -15.781   0.000
 BCOST         -0.435          0.025       -17.117   0.000
```

VOT= (BTIME/ BCOST)*60 = 12.34 euro/hour

```
+--------------------------------------------+
| Discrete choice (multinomial logit) model  |
| Maximum Likelihood Estimates               |
| Dependent variable              Choice     |
| Weighting variable                None     |
| Number of observations            2727     |
| Iterations completed                 5     |
| Log likelihood function       -1513.341    |
| R2=1-LogL/LogL*  Log-L fncn  R-sqrd  RsqAdj |
| No coefficients  -1890.212    0.199   0.197 |
| Constants only   -1803.763    0.161   0.158 |
| Chi-squared[ 8]           =    580.845     |
| Prob [ chi squared > value ] =    0.000    |
| Response data are given as ind. choice.    |
| Number of obs.=  2727, skipped   0 bad obs. |
+--------------------------------------------+
+---------+--------------+----------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+--------------+----------------+--------+---------+
 K              1.723          0.353         4.873    0.000
 BTIME         -0.098          0.006       -16.294    0.000
 BCOST         -0.477          0.027       -17.635    0.000
 BINCOME       -0.265          0.024       -11.196    0.000
 BAGE           0.015          0.004         3.591    0.000
 BWORK         -0.206          0.101        -2.040    0.041
 BEDU          -0.167          0.085        -1.960    0.050
 BFEMALE       -0.058          0.088        -0.657    0.511
 BFREQ         -0.165          0.053        -3.119    0.002
```

VOT= (BTIME/ BCOST)*60 = 12.34 euro/hour

Table 6. VOT estimation for some categories of respondents. Values in euro/hour.
Estimated model: $V = K + \beta_{time} time + \beta_{cost} \cos t$

	Full sample	**"Clean" sample**	Δ
Work trips	15.57	15.10	*-3.01%*

Non-work trips	11.07	9.96	*-10.03%*
Income lower than 20.000	9.96	9.32	*-10.73%*
Income higher than 20.000	18.57	17.80	*-4.15%*

Table 7. Logistic regression of inconsistence choices as a function of socio-economic variables in SP exercise (consistent choices=0, inconsistent choices=1).

VARIABLES	(A)	(B)
Age	0.031***	0.045***
	(0.010)	(0.013)
Female	0.035	-0.129
	(0.249)	(0.285)
Education	-0.183	-0.366
	(0.212)	(0.252)
Income	-0.020	-0.103
	(0.063)	(0.076)
Work trip	-0.497*	-0.488
	(0.280)	(0.309)
Freq. > 3 journeys	0.110	0.181
	(0.348)	(0.376)
Evaluation of travel alternatives (yes)	-0.048	-0.318
	(0.258)	(0.295)
Reason of the travel choice (1 if train is faster, cheaper or safer than other options)	0.456*	0.491*
	(0.247)	(0.283)
Perceived quality	-0.109*	-0.175**
	(0.059)	(0.072)
Economic value of delays	0.014	0.047
	(0.046)	(0.051)
Constant	-1.135	0.190
	(1.023)	(1.196)
Log-likelihood	*-217.47*	*-163.03*
Chi-squared	*29.80*	*42.45*
Observations	*407*	*279*

Figure 3. Bayesian Network.

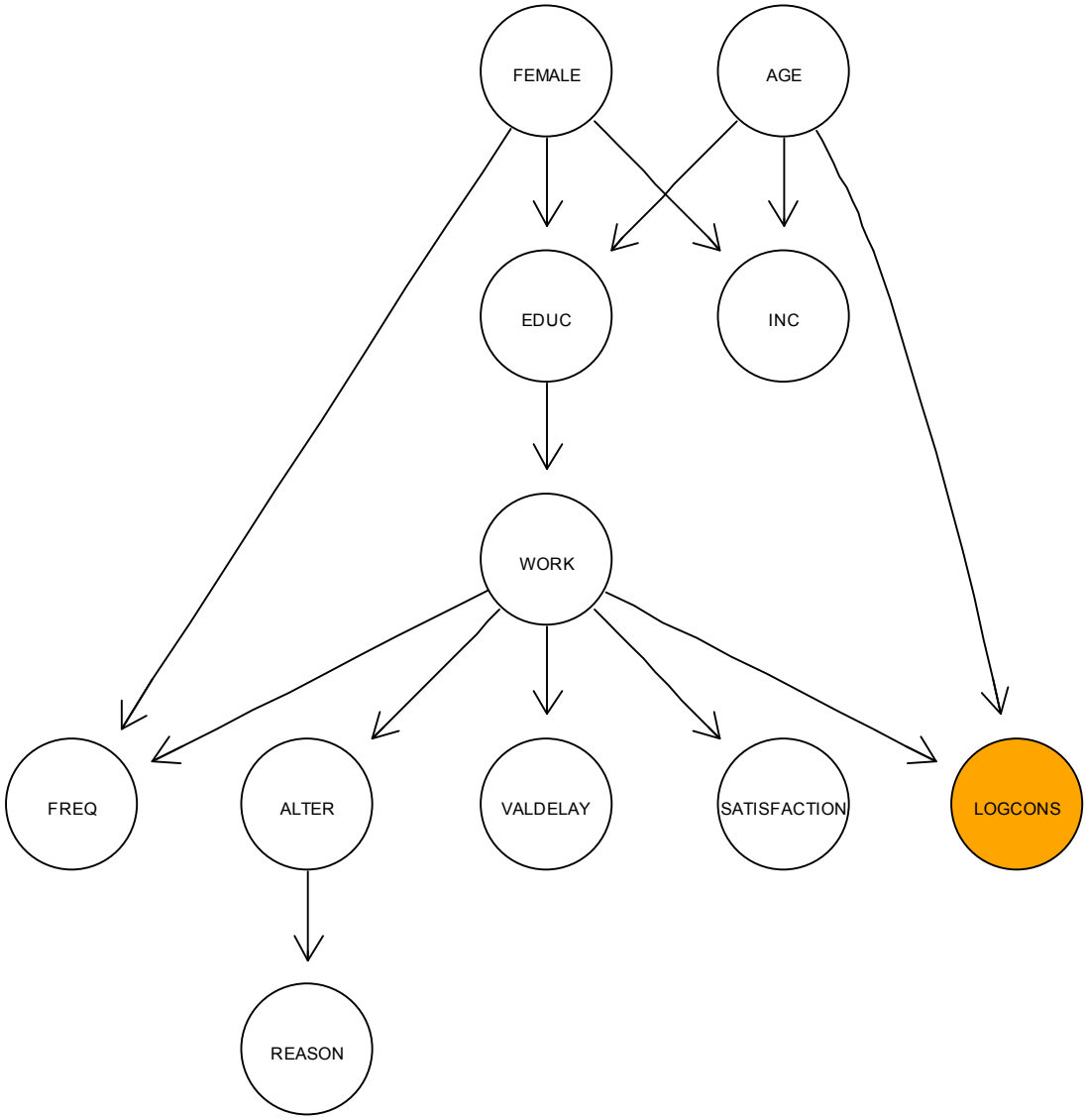


Table 8. Parameters of node LOGCONS. Conditional probability table.

Work = work				
Logical consistency	**Age**	**Age**	**Age**	**Age**
Consistent	0.875	0.773	0.800	0.846
Inconsistent	0.125	0.227	0.200	0.154
Work= nowork				
Logical consistency				
Consistent	0.714	0.822	0.722	0.404
Inconsistent	0.286	0.178	0.278	0.596

Table 9. Structure of the questionnaire and descriptive statistics.

		Absolute number	%
Gender (FEMALE)	Male	239	*58.72*
	Female	168	*41.28*
Reason for the trip (WORK)			
	Work trip	191	*46.93*
	Non work trip	216	*53.07*
How many times in a month do you travel on this rail line? (FREQ)			
	Less than one	177	*43.49*
	Between 1 and 3	158	*38.82*
	Between 3 and 5	41	*10.07*
	More than five	31	*7.62*
Age (AGE)			
	Less than 30	99	*24.32*
	Between 31 and 45	150	*36.86*
	Between 46 an 60	109	*26.78*
	Over 60	49	*12.04*
Education level (EDUC)			
	Elementary school	1	*0.25*
	Junior high school	27	*6.63*
	Senior high school	170	*41.77*
	University	209	*51.35*
Have you ever evaluated costs and benefits of other travel options (car, bus, etc.) for the trip between Turin and Milan? (ALTER)			
	Yes	177	*43.49*
	No	230	*56.51*
Why did you choose the railway? (REASON)			
	It is cheaper compared with other options	93	*22.85*
	It is faster compared with other options	17	*4.18*
	It is safer compared with other options	49	*12.04*
	I do not like to travel by car	52	*12.78*
	I do not have a car	71	*17.44*
	I like to travel by train / it is more comfortable compared with other options	125	*30.71*
Net yearly income (INC)			
	Less than 5.000	47	*11.55*
	Between 5.000 and 10.000	46	*11.30*
	Between 10.000 and 15.000	57	*14.00*
	Between 15.000 and 20.000	71	*17.44*
	Between 20.000 and 25.000	63	*15.48*
	Between 25.000 and 30.000	44	*10.81*
	Between 30.000 and 40.000	38	*9.34*
	Between 40.000 and 50.000	14	*3.44*
	Between 50.000 and 70.000	17	*4.18*
	More than 70.000	10	*2.46*

		Absolute number	%
How do you judge the quality of the service in terms of punctuality? (SATISFACTION)			
	1 (very bad)	4	*0.98*
	2	20	*4.91*
	3	33	*8.11*
	4	33	*8.11*
	5	47	*11.55*
	6	77	*18.92*
	7	84	*20.64*
	8	81	*19.90*
	9	16	*3.93*
	10 (excellent)	12	*2.95*
Do you consider train delays as an economic damage? (VALDELAY)	1 (severe damage)	2	*0.49*
	2	21	*5.16*
	3	43	*10.57*
	4	20	*4.91*
	5	42	*10.32*
	6	46	*11.30*
	7	27	*6.63*
	8	33	*8.11*
	9	39	*9.58*
	10 (not at all)	134	*32.92*
Have you ever tried to evaluate the damage in monetary terms?			
	Yes		
	No		

Stated preference exercise.

	Choice 1	
	Scenario 1	Scenario 2
Travel time	1h 30	1h 15
Price	8 euro	12.30 euro
	Choice 2	
	Scenario 1	Scenario 2
Travel time	1h 45	1h 15
Price	8 euro	12.30 euro
	Choice 3	
	Scenario 1	Scenario 2
Travel time	1h 30	1h 15
Price	8.50 euro	15.80 euro
	Choice 4	
	Scenario 1	Scenario 2
Travel time	1h 30	1h 15
Price	8.50 euro	13.50 euro
	Choice 5	
	Scenario 1	Scenario 2
Travel time	1h 40	1h 20
Price	8.50 euro	13.50 euro
	Choice 6	
	Scenario 1	Scenario 2
Travel time	1h 45	1h 15
Price	8.50 euro	15.80 euro
	Choice 7	
	Scenario 1	Scenario 2
Travel time	1h 45	1h 15
Price	8.50 euro	13.50 euro
	Choice 8	
	Scenario 1	Scenario 2
Travel time	1h 20	1h 40
Price	12.30 euro	8 euro
	Choice 9	
	Scenario 1	Scenario 2
Travel time	1h 40	1h 20
Price	8.50 euro	15.80 euro

References.

Ahmed F., Vaidya K.G., 2004, *The valuation of travel time savings in least developed countries: theoretical and empirical challenges and results from a field study*, WCTR '04.

Bates J., 1988, *Econometric issues in stated preference analysis*, Journal of Transportation Economics and Policy 22(1), 59–69.

Becker G.S., 1965, *A theory of the allocation of time*, The economic Journal, Vol.75, No. 299.

Bonsall P.W., Kirby H.R., 1984, *Urban appraisals*, Working Paper, Institute of Transport Studies, University of Leeds, Leeds, UK.

Bouffioux C., 2002, *Conjoint Analysis of Stated Preference*, Quality differences – SSTC project CP-TR-03, Bruxelles.

Cherchi E., 2003, *Il valore del tempo nella valutazione dei sistemi di trasporto*, Franco Angeli editore.

DeSerpa A., 1971, *A theory of the economics of time*, Economic journal, 81, pp. 828-846.

Domencich T.A., McFadden D., 1975, *Urban travel demand: a behavioural analysis*, Amsterdam, North-Holland

Evans A., 1972, *On the teory of the evaluation and allocation of time*, Scottish Journal of Political Economy, 19, pp 117.

Foster V., Mourato S., 2002, *Testing for consistency in contingent ranking experiments*, Journal of Environmental Economics and Management, Vol.44, pp.309-328.

Foster V., Mourato S., 1997, *Are consumers rational? Evidence from a contingent ranking experiment*, Paper presented at the 8th Annual Conference of the European Association of Environmental and Resource Economists, Tilburg, The Netherlands, 28-28 June.

Fowkes A.S., 2000, *Recent development in stated preferences techniques in transport research*, Stated Preference Modelling Techniques , pp.37-52 2000 (published by PTRC edited by J. de D Ortuzar).

Fowkes T., Wardman M., 1988, *The Design of Stated Preference Travel Choice Experiments*, Journal of Transport Economics and Policy, 22(1), pp. 27-44.

Fowkes T., Watson S.M., Toner J.P., Wardman M., 2000, *Efficiency properties of orthogonal stated preference designs*, Stated Preference Modelling Techniques , pp.91-101, (published by PTRC edited by J. de D Ortuzar).

Heckerman D., 1996, *A tutorial on learning with Bayesian Networks*, Microsoft Research, Technical Report MSR-TR-95-06.

Hensher D.A., 1994, *Stated preference analysis of travel choices: the state of practice*, Transportation 21 (2).

Hensher D.A., Rose J.M, Greene W.H., 2005, *Applied Choice Analysis. A Primer*, Cambridge University Press, UK.

Huber J., Zwerina K., 1996, *The Importance of Utility Balance in Efficient Choice Designs*, "Journal of Marketing Research", 33(3), pp. 307-317.

Jara-Diaz S., 1990, *Consumer's surplus and the value of travel time savings*, Transportation Research B, 24, pp73-87.

Johnson M., 1966, *Travel time and the price of leisure*, Western Economic Journal, 4, pp. 135-145.

Kroes E.P., Shelron R.J., 1988, *Stated Preference Methods: An Introduction*, Journal of Transport Economics and Policy, 22-1, Jan. 1988. pp.11-25.

Kuhfeld W.F., Tobias R.D., Garrat M., 1994, *Efficient Experimental Design with Marketing Research Applications*, Journal of Marketing Research, 31(4), pp. 545-557.

Layard R. and Glaister S., 2003, *Cost benefit analysis 3rd Edition*, Cambridge University Press.

Manski C., 1977, *The structure of random utility model*, Theory and decision, 8, pp. 229-254.

McFadden D., 1973, *Conditional logit analysis of qualitative choice behaviour*, Frontiers in econometrics, P. Zarembka ed., Acedemic Press, New York.

McFadden D., 1986, *The Choice Theory Approach to Market Research*, Marketing Science, Vol. 5, No. 4, Special Issue on Consumer Choice Models (Autumn, 1986), pp. 275-297.

Mullainathan S., Thaler R.H., 2000, *Behavioural Economics*, MIT department of Economics Working Paper n. 00-27.

OECD, 2006, *Cost-Benefit Analysis and the Environment: Recent Developments*, chapters 8 and 9, OECD Publications 2006.

Oort O., 1969, *The evaluation of travelling time*, Journal of transport economics and policy, 3, pp. 279-286.

Pearl J., 2000, *Causality: models, reasoning, and inference*, Cambridge University Press.

Saelensminde K., 2003, *The impact of choice inconsistencies in stated choice studies*, Environmental and resource economics, 23, pp. 403-420.

Scutari M., 2010, *Learning Bayesian Networks with the bnlearn R package*, Journal of Statistical Software, Vol. 35, n.3.

Small K., 1982, *Scheduling of consumer activities: work trips*, American Economic Review, 72, pp. 467-479.

Train K., Mc Fadden D., 1978, *The good/leisure tradeoff and disaggregate work trip mode choice models*, Transportation Research, 72 pp. 467-479.

Trt, 2001, *EVA-TREN - Improved Decision-Aid Methods and Tools to Support Evaluation of Investments for Transport and Energy Networks in Europe* .

UNITE, (2001), *Unification of accounts and marginal costs for transport efficiency*, Valuation conventions for UNITE.

Wardman M. and Waters W.G., 2001, *Advances in the evaluation of travel demand savings*, Transportation research, 37 E(2/3), pp.85-90.

Zwerina K. and Joel H., 1997, *Deriving Individual Preference Structures for Practical Choice Experiments*, Working Paper, Fuqua School of Business, Duke University.

Consumers' satisfaction and quality: evidence from a cross country comparison about railway transport

Abstract. In the last decades several surveys have been conducted in order to observe how consumers' satisfaction with Services of General Interest SGI) differs across EU countries. Many times survey results are used as surrogate for quality indicators: the more satisfied the users, the higher the quality of the service. Whether this procedure is appropriate or not represents the research question faced by this paper. In this work we investigate the connections between objective quality and self reported satisfaction with rail transport, using Eurobarometer surveys for 2000, 2002 and 2004. In the same model we include as regressors some individual characteristics of the respondents, in order to verify whether the evidence found in previous literature, which links satisfaction and socio-demographic features, is confirmed in our context. Moreover, beyond the size of the relationship, our purpose consists in understanding how individual characteristics affect perceptions.

Contents.

1. Introduction. 76
2. The data. 77
3. The empirical model. 79
4. Results. 81
5. Conclusions. 85
Notes 87
Appendix 1. 88
Appendix 2. 97
    The data. 97
    The empirical model. 98
    Results. 98
    Conclusions. 99
References. 105

1. Introduction.

In the last years the interest for consumers' experience with Services of General Interest (SGI) largely increased. Several surveys have been conducted in order to observe how consumers' satisfaction differs across EU countries. Many times the concepts of satisfaction and quality are used interchangeably: the more satisfied the users, the higher the quality of the service. But can we really consider them as a single construct? Some authors pointed out how, when agents are asked to express a judgement, several sources of bias are involved.

Attitudes are based on personal perceptions and evaluations which, in turn, depend on several unobservable and subjective factors. Some examples come from the health economic literature (McFadden *et al.*, 2005). Each person has a perception of her own health status, based on some objective parameters, which can significantly differ, *ceteris paribus*, from the status perceived by another individual.

The same reasoning applies when individuals are required to express a judgement about other aspects of their life, such as Services of General Interest (SGI). In particular, his work focuses on perceived satisfaction with railway transport. Of course, compared with the previous example about self-rated health status, when asked to evaluate their satisfaction with rail transport individuals are perfectly aware of the price they paid, of coaches' cleanliness and of all relevant features of the service. However, even in this case perception can be affected by individual-specific characteristics, such as income, gender, education, but also by group-specific features, social norms etc. (Bertrand and Mullainathan, 2001). The issue of interpersonal and cross-cultural incomparability in survey research has been identified and studied by many researchers, who developed new methods in order to overcome this problem (King and Wand, 2007).

Surveys results are often interpreted as quality benchmarks by policy makers and regulators. In the case of railways, poor evidence is available about objective quality of the service and surveys are commonly used in order to compare different services or countries. For example, the analysis of Eurobarometer Special Issues n.53 (2000), n.58 (2002) and n.62 (2004) provides us the picture represented in figure 1. Asked about satisfaction with price and quality, respondents from fourteen European countries expressed their opinion. The picture represents the departure from the mean of quality and price satisfaction in each country. The interpretation of this scenario usually leads to two conclusions. Taking Italy as reference, we could infer that:
1. in Italy railway transport quality is low;
2. transport quality is lower in Italy than in other countries (for instance Denmark or Greece).

The first statement assumes consumers' satisfaction to be a good predictor of objective quality. The second conclusion assumes that perceptions are not only good predictors of quality but that they are also fully comparable among different groups of individuals.

The research question pursued in the following lines concerns the verification of these two hypotheses and the investigations of the effects of individual and group-specific characteristics on perceived satisfaction.

Here we analyse the results from three Eurobarometer Issues (2000, 2002, and 2004), which include more than 26,000 interviews about satisfaction with rail transport in EU countries. In our model we include as predictors for satisfaction both individual characteristics and railways indicators. The latter should capture the most relevant differences in railway services across Europe.

A final section is devoted to a brief discussion about the problem of endogeneity of some among our regressors. A new model is estimated, using the results from another survey, conducted on the regional train between Turin and Milan in June and July 2008.

2. The data.

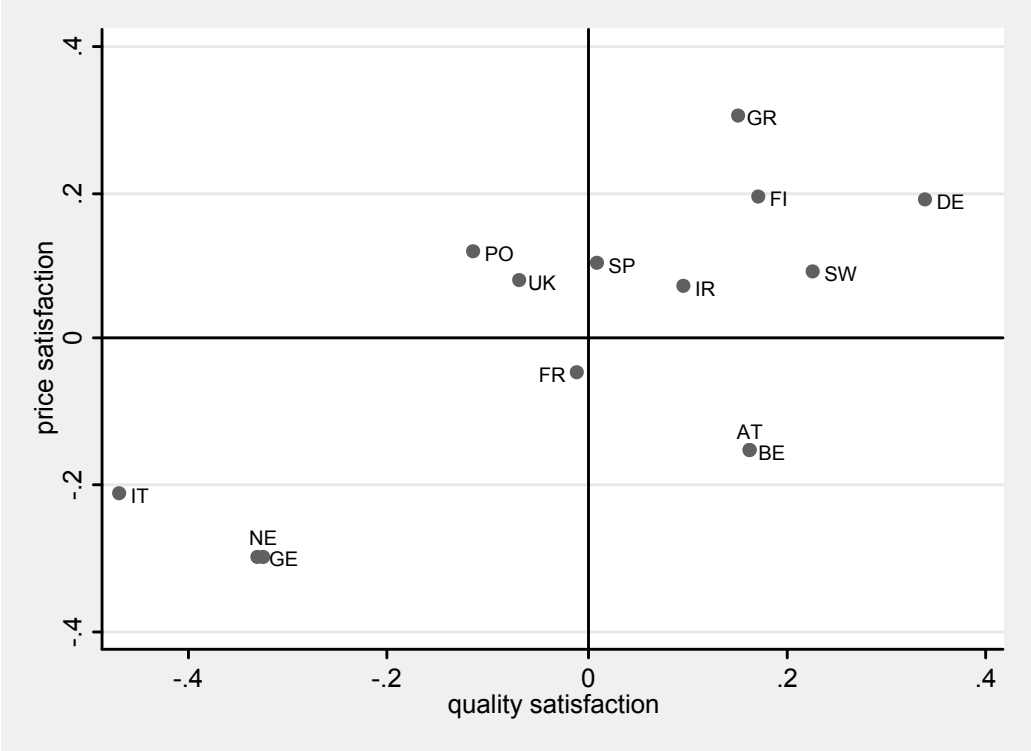
Eurobarometer surveys. Since 1973, the European Commission has been monitoring the evolution of public opinion in the Member States, thus helping the preparation of reports, and the evaluation of different policies. Eurobarometer surveys collect a large amount of information about individuals' actual attitudes, such as vote intention or media use, and their satisfaction with life in general and some specific issues, such as government policies or public services.

Satisfaction with SGI have been widely investigated: Eurobarometer Special Issues n. 47 in 1997, n. 53 in 2000, n.58 in 2002, n.62 in 2004 and n.63 in 2005 focus on the quality perceived by consumers about the supply of services such as gas, electricity, telecommunications and transportation. Here we focus on individuals' satisfaction with rail transport between cities, which has been studied in all Eurobarometer volumes summarised above.

In 1997 and 2005 surveys, question wording slightly differs compared with other Special Issues. Then, the analysis deals with 2000, 2002 and 2004 data, which are fully comparable[1]. Eurobarometer Special Issues n. 53, n. 58 and n.62 present one question about both price and quality satisfaction. In the first case, respondents are asked to express their opinion about prices, by choosing a category in a three point scale: three if they find the service to be "fair" priced, two when the price is judged as "unfair" and one if it is "excessive". Similarly, the quality of the rail transportation service can be defined according to the following scale: four if it is defined as "very good", two if it is just "fairly good", three if the quality is "fairly bad" and four for "very bad" quality. Figure 1 reports the average values for satisfaction with both price and quality, obtained by pooling the three samples. Countries in the low-left corner are characterised by lower average levels of satisfaction, while the opposite holds for the nations in the top-right one. Descriptive statistics about price and quality satisfaction are summarised in table 1.

Eurobarometer surveys provide us with a large amount of information about respondents' characteristics, such as age, gender, education, etc. Unfortunately some of these features are not available for all volumes. Then, we run our model separately on each dataset, in order to be able to include and investigate the effect of these issue-specific indicators.

Figure 1. Departures from the mean of price and quality satisfaction in EU countries.



Source: Eurobarometer Special Issues n.53 in 2000, n.58 in 2002, n.62 in 2004

Macroeconomic indicators. The inclusion of macroeconomic indicators is based on the assumption that the economic and social environment in which individuals act has an effect on their perception of satisfaction. For example, it seems reasonable to guess a negative relation between high unemployment rates and satisfaction. Other researchers (Fiorio et al., 2007) found significant results concerning GDP, GDP growth rate and employment rates. Usually macroeconomic indicators are country-specific. Since Eurobarometer surveys classify respondents based on their region of residence (at NUTS II level) we use as a source for macroeconomic indicators the Eurostat Regional Statistics database, in order to verify whether, compared with previous evidence, the inclusion of regional (instead of national) data improves the significance of our results.
Our predictors comprehend net disposable income of household expressed in PPP, the yearly change in disposable income, population density and unemployment rates.

Moreover, the suicide rate per 100.000 inhabitants (at NUTS II level as well) has been included in our models, based on the evidence found by Koivumaa-Honkanen et al. (2001) which links higher suicide rates with lower levels of satisfaction.

Railways indicators. As we said in the first paragraph, we add to individual characteristics and macroeconomic indicators seven regressors, representative for some features of railway transport which are supposed to influence individuals' preferences.

Following FitzRoy and Smith (1995), who studied the demand for rail transport in EU countries, we measure rail fares as passenger revenue per passenger-kilometer, expressed in euros and converted to a common currency using purchasing parity exchange rates. Descriptive statistics are reported in figure 2. Yearly change in fares measures the growth rate of the previous indicator.

Again, according to FitzRoy and Smith (1995), the demand for rail transport can be defined as passenger-kilometer per capita. Figure 5 in the appendix shows some evidence from EU countries.

The predictor "accidents" represents the number of fatalities in railway accidents, without any weight for countries' population. The underlying assumption is that country size does not matter since fatal accidents involving railways have a wide echo on newspapers and other media. Figure 4 reports the average number of passengers killed in railway accidents between 2000 and 2004. "Monopoly" is a dummy variable equal to one if a unique rail transport provider operates in the country and equal to zero if in the market acts more than one provider.
All these variables are at country level, as regional data are not available. In most cases our source is the World Bank's Railways Database, other sources are summarised with more details in table 7.

Finally, we built two predictors for the supply of rail transport and motorways, which are railways' main competitor. In this case data are available for NUTS II regions and the regressors are defined as the ratio between the length of railways (or motorways) and the region area. Figure 3 reports some data about the supply of railways and motorways at country level. In general, the former have a higher weight with respect to motorways in less developed countries.

3. The empirical model.

Since respondents have been asked to state their satisfaction based on a three (in the case of price) and four (for quality) points scale, answers are only ordinally comparable. In other words we do not know what the relative distance between satisfaction answers is, but we assume that all individuals share the same interpretation of each possible answer.

The true level of satisfaction can be represented as a continuous, unobserved latent variable Y^*, which is equal to:

$$Y^* = x\beta + e$$

where the random disturbance term has a logistic distribution. Instead of Y^* we observe Y, our ordinal variable. We can think at Y as a collapsed version of Y^*, whose value depends on whether or not the continuous variable Y^* crossed a particular, unknown, threshold ($\alpha_1 < \ldots < \alpha_j \ldots < \alpha_M$):

$$Y_i = 0 \quad \text{if} \quad Y_i^* \le \alpha_1$$
$$Y_i = 1 \quad \text{if} \quad \alpha_1 \le Y_i^* \le \alpha_2$$
$$\vdots$$
$$Y_i = M \quad \text{if} \quad Y_i^* > \alpha_M$$

where M is the number of categories of the ordinal dependent variable.
Then, we can estimate the probability that Y will take on any particular value through standard ordered logistic regression:

$$P(Y_i > j) = g(x\beta) = \frac{\exp(\alpha_j + x_i\beta)}{1 + \left[\exp(\alpha_j + x_i\beta)\right]}, j = 1,2,...,M-1 \qquad (1)$$

Note that ordered logit models are multi-equational models: they correspond to a series of binary logistic regressions where categories are combined. For example, if we want to analyse satisfaction with prices, which is coded as a categorical variable with three outcomes, two separate equations are estimated: the first one compares one category (e.g. the most positive judgements) with a combination of the other two, while the second equation contrasts categories 1 and 2 versus the third one. As a consequence, we get for each predictor a number of coefficients equal to $M-1$.

However the model (1) is based on the assumption that the relationship between each pair of outcome groups is the same. In other words, the proportional odds assumption forces the coefficients to be the same. Since this hypothesis is often violated, we test in our models its validity.
When the assumption is rejected, we estimate another model, by relaxing the proportional odds hypothesis for a subset of predictors. As a first step we identify those regressors for which the assumption holds and, as a second step, we estimate a new model where all other predictors' coefficients are allowed to change for each pair of outcome groups.
Following Williams (2007) we run a new model which is a hybrid between the one described by (1) and the generalised ordered logit model:

$$P(Y_i > j) = g(x\beta_j) = \frac{\exp(\alpha_j + x_i\beta_j)}{1 + \left[\exp(\alpha_j + x_i\beta_j)\right]}, j = 1,2,...,M-1 \quad (2)$$

where the Betas are not the same for each value of j.

As controls we use a set of individual characteristics. Some of them (gender, age, education, marital status, occupational group, political views) are available for all Eurobarometer Special Issues, while some questions have been asked only in some Eurobarometer waves. Since it is interesting to investigate the relation between this second group of variables (such as life quality, income, community size) and satisfaction we replicate our analysis on each sample. Moreover, this procedure allows us to compare patterns of satisfaction between EU Old Member States (OMS) and New Member States (NMS), which have been included in the survey only from 2004. In this case we estimate two models: in the first one we take as reference the group of OMS and we include country dummies for all NMS, while in the second model we do the opposite.

Finally, we pool together the three samples in order to check the consistency of previous results and to include all country dummies for OMS, taking Italy as reference.

4. Results.

The presentation of the results is organised as follows. Tables 2-6 in the appendix report the estimations for each Eurobarometer survey (2000, 2002 and 2004) and the results from the pooled sample (table 5 and 6). Here some evidence is summarised, pointing out the main differences among the estimated models.

Individual characteristics. Respondents' *gender* seems to affect satisfaction, especially with prices. Females tend to be less satisfied than males, and this result is consistent across all samples, with the exception of 2000. Considering the pooled sample and holding all other variables constant, the odds of having a more negative opinion about ticket prices are 1.14 times larger for women than men.

The effect of *age* on satisfaction with both price and quality of the service is significant and consistent across all samples. The sign is positive, which means that elderly are more likely to be satisfied compared with younger people, even if the dimension of the coefficient is rather small. If we consider the pooled sample, the coefficient of age on price satisfaction is about 0.008. Translated in terms of changes in probabilities (not reported in table 6), this result shows how, shifting from the minimum value of age (15 years) to the maximum (99), the probability of defining the fares as "excessive" reduces of about 4 per cent, while the probability of finding them "fair" increases of about 15 per cent.

Regressors about respondents' *occupation* do not show clear and strong results even if we could have assumed some effects on satisfaction with train transport. For example, one would expect a positive relation between price satisfaction and the dummies for students and retired people, since reduced price tickets are often available for these two categories of travellers. This hypothesis is not strongly supported by the evidence. Students seem to be more satisfied with prices in 2000 and 2002, while the result is not confirmed in 2004. The coefficient about quality satisfaction is positive as well, but weakly significant. Unexpectedly, the strongest result for retired people concerns their feelings about quality. They seem to be more likely to be satisfied with this feature of the service but, compared with the other occupational groups, they apparently do not judge more generously the level of fares (with the exception of 2000). Evidence arising from the inclusion of dummy variables for the other professional categories is not marked by consistency across the samples.

Political views apparently matter, as center and right voters tend to be more satisfied with both price and quality. Results become more significant and coefficients' dimension is larger when data do not allow us to control for income at individual level (tables 2 and 3 vs tables 4 and 5).

As we could expect, those who personally made a *complaint* in the last two months, either to a complaint-handling body or to the service provider are less likely to give positive feedbacks about their satisfaction. Referring to the pooled sample, the probability of defining the price as "fair" decreases of about 15.6 per cent for the travellers who made a complaint as opposed to those who did not. The same holds for quality satisfaction, even if this time the proportional odds assumption is rejected. The generalised ordered logit estimation shows (table 6) that the effect of complaining is particularly large for the second outcome category, which is the probability of finding the quality "fairly bad".

Keeping all other variables constant, respondents who finished full-time education up to fourteen years (*low education*) seem to be less satisfied than the others, at least in 2004. This result is weakly confirmed for 2002, while it is not statistically significant in 2000. The same reasoning applies to *separated and divorced* people.

Predictors not constant across all surveys. As we said before, some individual characteristics are not available for all Eurobarometer issues but they could be relevant in order to understand changes in our dependent variable. Respondents have been asked about their *life satisfaction* in 2000 and 2004. The rationale behind the inclusion of this variable is that people satisfied with their life are more likely to be satisfied also with other collateral components of their existence, such as railways transport. Even if the question wording slightly differs between the two surveys[2], the evidence is very similar and it supports our assumption. In our models the regressor for life satisfaction is a dummy variable equal to one for those respondents who defined themselves "fairly" or

"very satisfied" with their life[3]. Results in tables 1 and 4 show how the positive relation between life satisfaction and the probability of being pleased with railway transport is relevant especially for prices. Switching from unsatisfied interviewees to satisfied ones, the probability to define fares as "fair" increases of about 11 per cent.

Expectations about *future life quality* (available only for 2004[4]) matter as well: optimism is related with positive opinions about both price and quality of the service.

Two questions, in 2000 and 2002 surveys, investigate individual wealth asking each respondent to choose his category of *income* out of a list and to state whether he is the person who contributes most to the household income or not. Two variables in our models try to capture any effect linked to individual wealth. Weak evidence can be observed in 2000. Respondents have been split into income quartiles: shifting from the lowest to the highest one, the probability of defining ticket prices as "fair" increases of about 4.3 per cent. However, any conclusion has to be based on the presumption that people do not lie when asked to reveal their income, which sounds quite optimistic.

The *size of the community* is significant in 2002 but not in 2004. A dummy variable equal to one identifies those respondents who live in a big city: apparently they tend to be more satisfied with quality than the others, which seems reasonable since in large towns transport networks are more developed compared with rural areas. This interpretation leads us to other two variables which concern the accessibility of the service. Those who declared to have *difficult access* to rail transport (this variable is available for all surveys) are definitely less likely to be satisfied compared with the respondents who have easy access. In particular, evidence is relevant for quality satisfaction, which makes sense because accessibility can be viewed as a component of service quality. Keeping all other variables constant to their mean (table 6), the probability of being "very satisfied" with quality decreases of about 19 per cent for travellers who have difficult access. The same evidence applies for those who do not have access at all. The variable *no access* is not available for 2004 since for that Eurobarometer issue only travellers who have access to the service have been asked about their satisfaction. As can be seen for the generalised ordered logit results in table 6, the effect of having no access is more extreme on lower levels of satisfaction.

Instead of investigating differences between travellers who have access to rail transport and those who do not, 2004 data allow us to distinguish between respondents who *use* the train and those who do not. The result is quite trivial: people who travel by train are more likely to be satisfied with prices[5], but it seems reasonable to assume that those who prefer other means of transport do not find particularly convenient to use railways.

Macroeconomic indicators. The first regressor analyses the relationship between satisfaction and the average *disposable income* at NUTS II level. Evidence is consistent across all surveys and it suggests a negative effect of the predictor on our dependent variables. This could sound somehow counterintuitive, as one could expect a positive

connection between wealth and satisfaction, as the one (weakly) observed for the income quartile. However, since the wealth indicator reported here refers to regions rather than individuals, the negative coefficient can be explained by the hypothesis that richest regions are characterised by more developed transport networks. Then, railways are in competition with many other travel options and, as a consequence, travellers' judgements are more severe. *Yearly change in disposable income* seems to be related to lower levels of satisfaction as well, but results are not significant in a couple of cases (2002, 2004).

Population density does not have any influence on the perception of service quality and price fairness.

Respondents who live in regions characterised by higher *unemployment* rates tend to be less satisfied with prices, while the effect on quality satisfaction is not confirmed across all Eurobarometer Issues.

Finally, contradictory results are shown for the *suicide* rate. Based on the literature suggesting a negative relation between life satisfaction and suicides, we would expect a negative coefficient. However, this hypothesis is not verified, and suicide rates seem to be linked to less extremely negative patterns of satisfaction with quality.

Railways indicators. An increase in *fares* is likely to reduce perceived satisfaction with both price and quality. This effect is consistent across all surveys. In principle, we would expect a positive relation between fares and quality satisfaction. However, results show the opposite relation: the higher the fares the lower the satisfaction with quality. In general, the evidence suggests the occurrence of endogeneity, since prices depend on several unobserved factors. Moreover, agents expressed their satisfaction based on the price they actually paid, while in our models we are controlling for an average measure for fares. This topic is discussed with more details in the second appendix.

Yearly change in fares is associated with lower levels of satisfaction, especially with prices.

The number of *fatalities* (not available in 2004) does not seem to be a powerful predictor for satisfaction. We would expect a connection with quality satisfaction but no evidence confirms this hypothesis.

Concerning market structure, m*onopoly* is associated with higher levels of satisfaction with prices. This result is consistent across all surveys, while the same evidence does not apply for quality satisfaction.

The *demand* for rail transport, defined as per capita passengers-kilometres, is supposed to be positively related to consumers' satisfaction: if travellers choose this mean of

transport probably they do like it. However, results show negative coefficients in all models but in the pooled sample, where we included country dummies.
The same conclusions apply for the railways and motorways length. Coefficients fluctuate between positive and negative values and any strong deduction can be drawn by the inclusion of these regressors in our model.

Country dummies. Country dummies have been included in our models in 2004 and in the pooled sample.

As we said, 2004 Eurobarometer Issue allows us to compare NMS with OMS. If we include in our model a dummy variable equal to 1 for NMS instead of country dummies the results (not reported in the tables) show that, compared with OMS inhabitants, respondents from NMS tend to be less satisfied with both price (coeff. about -0.47) and quality (coeff. about -0.94).

Output reported in table 4 allows us to discriminate among NMS. Hungary, Poland and Slovakia are the countries characterised by the lowest coefficients for price satisfaction. Moreover, the generalised ordered logit estimation warns us about the weakness of our results, as the proportional odds assumption is rejected for most of country dummies. With the exceptions of Lithuania and Slovenia, NMS respondents tend to be more at the lower extreme of price satisfaction (i.e. "excessive") than OMS people are. The same reasoning holds for quality satisfaction, even if this time the proportional odds assumption is violated only in two cases.

Pooled sample outcomes, reported in table 6, allow us to analyse the group of OMS taking Italy as reference. Again, we can observe how the ordered logit model does not represent the best option in order to properly capture country effects. Compared to our country, the probability of being satisfied with prices seems to be higher in all other nations but Netherlands, Germany and Austria. People from some countries (Spain, Portugal, Finland, Sweden and Greece) tend to express less extreme judgements for dissatisfaction.
Concerning quality, all countries' coefficients are positive, which means that non-italian respondents are more likely to report higher levels of satisfaction. As showed by the generalised ordered logit results, this time differences do not focus on the lowest category but on the higher one. In several countries (Danemark, Ireland, Finland, Sweden and Austria) satisfied respondents tend to state their satisfaction more generously[6] compared with Italians.

5. Conclusions.

Results presented in the previous section allow us to draw some conclusions for each group of predictors.

Considering *individual characteristics*, the evidence depicted in our models is consistent with other literature findings. Fiorio and Florio (2007) focused on consumers' satisfaction with electricity providers across EU countries. They found negative connections between satisfaction, respondent's gender (female) and political views (left voters). These results have been confirmed in our work, as well as for the weak link between some occupational groups (students, retired) and satisfaction. Other researchers (Hall and Dornan, 1990) found robust evidence linking age and satisfaction with healthcare services: older patients tend to be more satisfied, as in our analysis.
Compared with other studies we have been able to include individual life satisfaction, which appeared to be one of the most significant predictors for satisfaction with rail transport services.

Macroeconomic indicators are in general not strongly significant with two exceptions: the disposable income and the unemployment rate. The result concerning income is supported by Fiorio and Florio (2007), even if in their study macroeconomic indicators were defined at country level. Apparently, the use of regional data does not improve findings' significance. The robustness of these results reduces when we enlarge the sample to NMS (2004).

The inclusion of *railways indicators* provided strong results for the level of fares and the structure of the market (monopoly), while in some cases the evidence is somehow contradictory (transport networks length) and in some others it is quite weak (demand).
We have to underline how some relevant features of railway transport have not been included in our model due to unavailability of data. Delays, cleanliness, coaches' comfort do matter in determining consumers' satisfaction. Moreover, our regressor for fares captures the relative price of train transport across countries but not across means of transportation. For example, motorways are tolled in some nations and free of charge in some others. For the latter group, you would expect a negative relation between fares and price satisfaction, as the opportunity cost for train transport is higher. However, the inclusion of a dummy equal to one for those countries without charged motorways and equal to zero for the others did not produce any significant result[7]. This issue is discussed in the second appendix, where the same experiment has been replicated on a sample of respondents who shared the same travel experience.

Recalling the two hypotheses suggested in the introduction, we can conclude that satisfaction is strongly influenced by individual characteristics. Moreover, the connection between satisfaction and objective quality seems to be affected by individual specific features such as age, gender or occupation. Concerning the comparability of patterns of satisfaction among different countries, results show how, after controlling for individual characteristics and socio-economic indicators, relevant differences characterise each country. These differences involve not only the absolute values of satisfaction, but also the way of expressing opinions, which appears more extreme in some nations compared with others. Since the country dummies include all the unobserved features for which we did not account in the model, this is just a conjecture.

However, it is an assumption which merits further research and finds theoretical and empirical verification in some recent literature about surveys (King and Wand, 2007).

Notes

[1] In 2004 the question wording about price satisfaction differs compared with the other two. However, answers can be rearranged in order to be consistent with the other datasets. More in detail, in 2004 respondents answered two questions about their satisfaction with prices. The first one was: "in general, would you say that the price you pay for rail service is affordable, not affordable or excessive?". The second question was: "and in general, would you say that the price you pay for rail services is justified or not?". We rearranged the answers by coding "3" (corresponding to "fair" in 2000 and 2002) for those who answered "justified" in the second question, "2" ("unfair" in the other issues) for those who said "not justified" in the second question but did not choose "excessive" in the first one. Consequently, we marked with "1" those who find the price "not justified" in the second question and "excessive" in the first one.
[2] In 2000 the question was: On the whole, are you very satisfied, fairly satisfied, not very satisfied or not at all satisfied with the life you lead ? In 2004 the question changed as follows: How would you judge the current situation in each of the following domains (life quality): very good, fairly good, fairly bad or very bad?
[3] In 2004 those who answered "very good" and "fairly good".
[4] The variable "future life satisfaction" recodes the answers at the following question: according to you, in five years, will the situation in each of the following domains (life quality) be much better, somehow better, identical, somehow worse, much worse than it is now? Our regressor is a dummy equal to one for those who answered "much better" or "somehow better" and for the respondents who said "identical" but answered "very good" or "fairly good" in the previous question about the quality of their life.
[5] Only travellers who have access *and* use the train have been asked about quality satisfaction.
[6] They define themselves "very satisfied" rather than "fairly satisfied".
[7] We dropped the dummy from the models presented here due to issues of statistical fit.

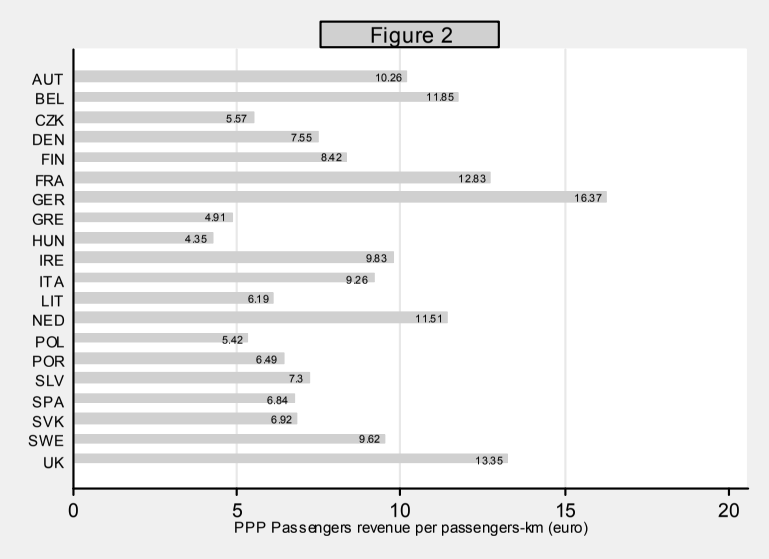
Appendix 1.

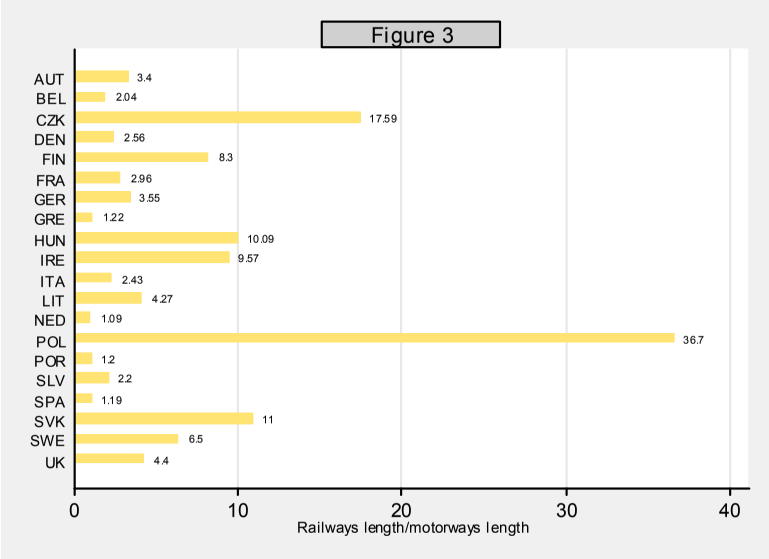
Table 1. Price and quality satisfaction across EU countries (pooled sample).

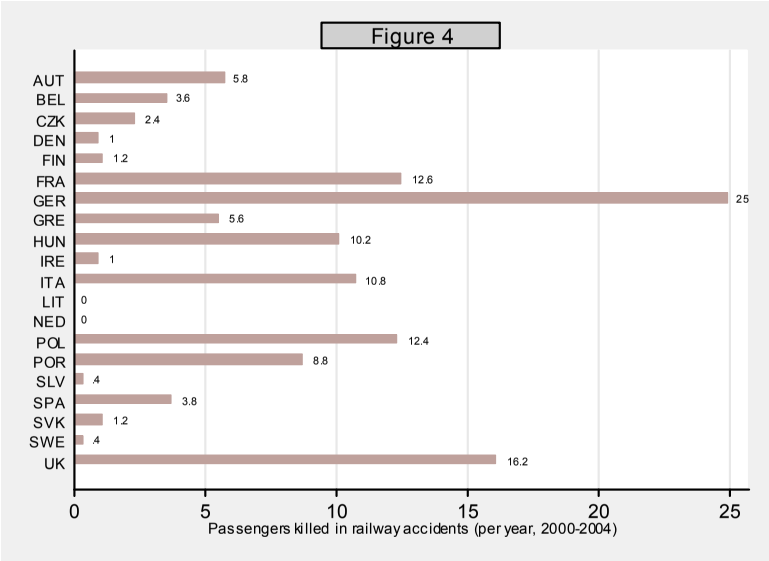
			Price satisfaction		
	Observations	Mean	Std. Deviation	Minimum	Maximum
Belgium	2,113	2.283	0.760	1	3
France	1,967	2.389	0.702	1	3
Netherlands	2,527	2.135	0.723	1	3
Germany	4,248	2.134	0.669	1	3
Italy	2,184	2.222	0.735	1	3
Denmark	2,329	2.623	0.553	1	3
Ireland	1,972	2.504	0.689	1	3
United Kingdom	2,833	2.515	0.624	1	3
Greece	1,995	2.736	0.482	1	3
Spain	2,231	2.538	0.599	1	3
Portugal	1,789	2.553	0.569	1	3
Finland	1,968	2.626	0.507	1	3
Sweden	2,416	2.524	0.562	1	3
Austria	2,113	2.283	0.760	1	3
			Quality satisfaction		
	Observations	Mean	Std. Deviation	Minimum	Maximum
Belgium	1,910	3.117	0.774	1	4
France	1,729	2.945	0.705	1	4
Netherlands	2,093	2.630	0.799	1	4
Germany	3,615	2.624	0.781	1	4
Italy	1,801	2.486	0.761	1	4
Denmark	2,097	3.296	0.732	1	4
Ireland	1,911	3.052	0.870	1	4
United Kingdom	2,515	2.886	0.857	1	4
Greece	1,746	3.105	0.687	1	4
Spain	1,853	2.965	0.644	1	4
Portugal	1,540	2.842	0.611	1	4
Finland	1,753	3.127	0.649	1	4
Sweden	2,241	3.183	0.763	1	4
Austria	1,910	3.117	0.774	1	4

Price satisfaction is ranked in three categories: 1 (excessive), 2 (unfair) and 3 (fair).
Quality satisfaction is classified in four classes: 1 very (unsatisfied), 2 (fairly unsatisfied), 3 (fairly satisfied) and 4 (very satisfied).

Figures 2-5. Descriptive statistics about railways in EU countries. Fares (figure 2), network length (figure 3), accidents (figure 4) and demand (figure 5). Year: 2000, 2002 and 2004 average values. Sources are summarised in table 7.







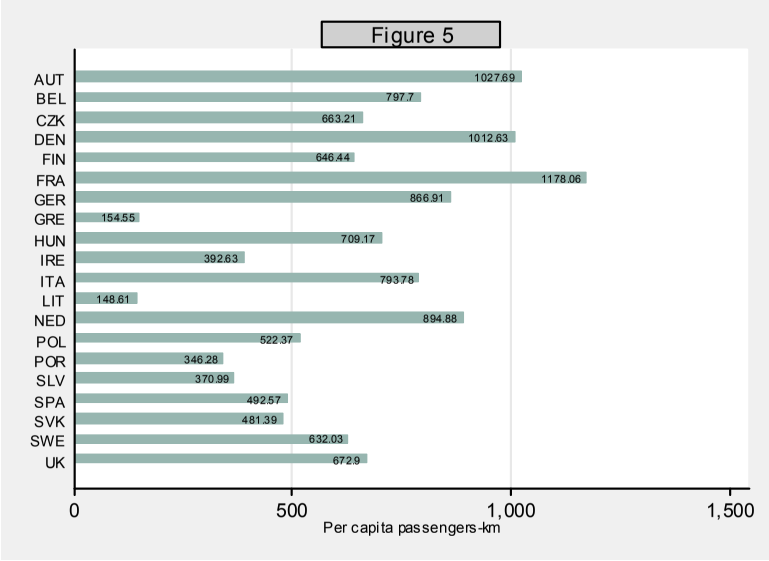


Table 2. Satisfaction with railway transportation services, generalised ordered logit reults (coefficients): 2000.

VARIABLES	PRICE SATISFACTION		QUALITY SATISFACTION		
	Generalized ordered logit		Generalized ordered logit		
	1	2	1	2	3
			Individual characteristics		
Female	-0.074		0.049		
Age	0.009***		0.007***		
Life satisfaction •	0.432***		0.288***		
Low education	-0.097		-0.178	0.006	-0.220**
Separated/divorced	-0.049		0.073		
Complaint	-0.599***		-0.953***		
Occupational groups:					
Student	0.321**		0.150		
Retired	0.325**		0.056		
Manual worker	0.168		0.059		
Manager	0.091		-0.079		
White collar	0.088		-0.114		
House person	0.021	0.504***	0.203		
Unemployed	0.151		0.145		
Political views:					
Center	0.010	0.160***	0.092		
Right	0.039		0.038		
DK/NA	0.085		0.109		
Income contributor •	0.035		0.056		
Income quartile •	0.052*		-0.026		
Difficult access	-0.696***	-1.362***	-1.942***	-2.092***	-1.649***
No access •	-2.034***		-3.225***	-2.012***	-0.930**
			Macroeconomic indicators		
Disposable income	-0.059**	-0.005	-0.087**		
Yearly change in disposable income	-0.076**	-0.095***	-0.043	-0.021	-0.051*
Population density	0.000*		0.000		
Unemployment	-0.018	-0.014	-0.017	-0.035**	-0.082***
Suicides	0.049***	0.017**	0.057***		
			Railways indicators		
Fares	-0.121***		-0.115***	-0.126***	-0.065*
Yearly change in fares	0.024***		0.022**	0.018***	0.006
Accidents	0.026*	-0.003	0.002		
Monopoly	0.370***	0.090	-0.745***		
Demand	-0.001***		-0.000		
Railways	-0.015***		-0.003		
Motorways	-0.019***		-0.019***		
Constant	4.936***	2.043***	5.536***	3.526***	0.882*
Log-likelihood	-6,810.844			-8,293.762	
Wald-chi squared	909.26			1,191.67	
Observations	7,973			8,074	

*** p<0.01, ** p<0.05, * p<0.1

Omitted categories are: "manual worker" (occupational groups), "left" (political views), "easy access".

• Variables not constant across all surveys.

Table 3. Satisfaction with railway transportation services, generalised ordered logit reults (coefficients): 2002.

2002 VARIABLES	PRICE SATISFACTION Generalized ordered logit		QUALITY SATISFACTION Generalized ordered logit		
	1	2	1	2	3
			Individual characteristics		
Female	-0.177***		0.085		
Age	0.009***		0.005***		
Low education	-0.196*		0.070		
Separated/divorced	0.100		0.082		
Complaint	-0.708***		-0.849***	-1.211***	-0.357
Community (big city) •	-0.083	0.005	0.218**		
Occupational groups:					
Student	0.235*		0.227		
Retired	0.164		0.223	0.263**	0.401***
Manual worker	-0.080		0.168		
Manager	0.014		0.212	-0.134	0.114
White collar	-0.043		0.238	-0.024	0.145
House person	-0.049		0.234**		
Unemployed	-0.066		-0.009	-0.030	0.064
Political views:					
Center	0.175***		-0.112	0.055	0.244***
Right	0.178***		-0.132	0.033	0.275***
DK/NA	0.015		0.052		
Income contributor •	0.179	-0.098	0.033		
Income quartile •	0.092**	-0.026	0.028		
Difficult access	-0.801***	-1.282***	-2.049***		
No access •	-1.500***		-2.852***	-2.072***	-1.113***
			Macroeconomic indicators		
Disposable income	-0.022***		0.003		
Yearly change in disposable income	0.061**	-0.018	-0.075***		
Population density	-0.000*		-0.000***		
Unemployment	-0.056***	-0.058***	0.021		
Suicides	0.023**		0.050**	0.024**	-0.036**
			Railways indicators		
Fares	-0.061**		-0.101**	-0.138***	-0.084***
Yearly change in fares	-0.002		-0.000	-0.014***	-0.032***
Accidents	0.019**	0.002	0.018	0.012	-0.006
Monopoly	0.920***	0.560***	-0.082	0.301	0.584***
Demand	-0.002***	-0.001***	-0.000	0.000	0.002***
Railways	0.014**		0.011**		
Motorways	-0.004		-0.006		
Constant	4.106***	1.615***	3.568***	1.939***	-1.868***
Log pseudo-likelihood	-5,843.865			-7,169.615	
Wald-chi squared	762.01			2,177.75	
Observations	6,926			7,075	

*** p<0.01, ** p<0.05, * p<0.1

Omitted categories are: “manual worker” (occupational groups), “left” (political views), “easy access”.

• Variables not constant across all surveys.

Table 4. Satisfaction with railway transportation services, generalised ordered logit reults (coefficients): 2004.

2004 VARIABLES	PRICE SATISFACTION		QUALITY SATISFACTION		
	Generalized ordered logit		Generalized ordered logit		
	1	2	1	2	3
			Individual characteristics		
Female	-0.125***		-0.100**		
Age	0.007***		0.009***		
Low education	-0.090	-0.134**	0.172*		
Separated/divorced	-0.177**		-0.064		
Complaint	-0.621***		-1.270***	-1.095***	-0.455***
Community (big city) •	-0.029		0.031		
Occupational groups:					
Student	-0.066		0.140		
Retired	0.045		0.414***		
Manual worker	0.058		0.300**		
Manager	-0.171		-0.078		
White collar	-0.072		0.084		
House person	-0.038		0.301**		
Unemployed	-0.346***		0.150		
Political views:					
Center	0.137***		0.147**		
Right	0.129**		0.180***		
DK/NA	0.032		0.011		
Life satisfaction •	0.470***		0.233***		
Future life satisfaction •	0.048	0.223***	0.179**		
Difficult access	-0.346***	-0.662***	-1.272***		
Use •	0.139	0.291***			
			Macroeconomic indicators		
Disposable income	-0.014		-0.114*	0.001	-0.055
Yearly change in disposable income	0.031*	0.004	0.007		
Population density	0.000		0.000		
Unemployment	0.017		0.003	0.006	-0.075***
Suicides	0.058***	0.035***	0.045***		
			Railways indicators		
Fares	-0.285***	-0.018	-0.062	-0.122***	-0.070*
Yearly change in fares	-0.062***	-0.019*	0.053**	0.004	-0.032**
Accidents	-0.148	-0.370***	-0.002		
Monopoly	-0.172*	0.158***	0.235*		
Demand	-0.000	-0.035***	0.001		
Railways	0.002		-0.013***		
Motorways	0.018**		-0.013		
			New Member States		
Czech Republic	-3.042***	-0.729***	-1.170***		
Hungary	-3.964***	-0.650*	-1.957***		
Lithuania	-0.729		-1.117**		
Poland	-2.930***	-0.788***	-2.141***	-1.417***	-0.807**
Slovakia	-3.885***	-1.613***	0.035	-0.968***	-0.297
Slovenia	-0.708**	0.615***	0.053		
Constant	6.103***	0.592	3.962***	1.237**	-0.693
Log pseudo-likelihood	-10,107.401			-6,356.996	
Wald-chi squared	1,474.35			8,955.41	
Observations	11,897			6,479	

*** p<0.01, ** p<0.05, * p<0.1

Omitted categories are: “manual worker” (occupational groups), “left” (political views), “easy access”.

• Variables not constant across all surveys.

Table 5. Satisfaction with railway transportation services, generalised ordered logit reults (coefficients): pooled sample.

VARIABLES	PRICE SATISFACTION		QUALITY SATISFACTION		
	Generalized ordered logit		Generalized ordered logit		
	1	2	1	2	3
			Individual characteristics		
Female	-0.137***		0.025		
Age	0.008***		0.013***	0.005***	0.005***
Low education	-0.130**		-0.005	0.127**	0.108*
Separated/divorced	-0.127***		0.174	0.022	-0.088
Complaint	-0.635***		-0.998***	-1.201***	-0.566***
Occupational groups:					
Student	0.094		0.093		
Retired	0.141**		-0.064	0.202***	0.240***
Manual worker	-0.038		0.080		
Manager	-0.079		-0.067	-0.183**	-0.012
White collar	-0.071		-0.048		
House person	0.095		0.025	0.204**	0.266***
Unemployed	-0.204***		0.079		
Political views:					
Center	0.192***		0.130***		
Right	0.119***		0.050	0.039	0.114**
DK/NA	-0.018	0.086*	0.140*	0.177***	0.042
Difficult access	-0.611***	-1.198***	-1.922***	-1.916***	-1.493***
No access	-1.368***		-2.917***	-1.934***	-1.227***
Year 2002	0.488***	-0.194**	-0.292**	-0.080	-0.180**
Year 2004	0.501***	-0.089	0.100	0.071	-0.222*
			Macroeconomic indicators		
Disposable income	-0.048**		-0.029		
Yearly change in disposable income	-0.001	-0.037***	-0.049***	-0.017	-0.008
Population density	-0.000		-0.000		
Unemployment	-0.024**		-0.013		
Suicides	-0.007		0.032***	0.009	-0.011
			Railways indicators		
Fares	-0.010	-0.092***	-0.050**		
Yearly change in fares	-0.006**		0.005		
Accidents	0.015*	-0.000	-0.007	-0.009	0.002
Monopoly	0.131***		0.046		
Demand	0.001	0.000	0.001		
Railways	-0.003		0.006**		
Motorways	-0.001		-0.000		
Country dummies	Yes		Yes		
Constant	1.197	-0.128	1.665*	-0.279	-3.450***
Log pseudo-likelihood	-22,591.226		-27391.931		
Wald-chi squared	5,267.21		13173.92		
Observations	26,684		27,296		

*** p<0.01, ** p<0.05, * p<0.1. Omitted categories are: “self employed” (occupational groups), “left” (political views), “easy access”.

Table 6. Country dummies in the pooled sample.

COUNTRIES	PRICE SATISFACTION		QUALITY SATISFACTION		
	Generalized ordered logit		Generalized ordered logit		
	1	2	1	2	3
France	-0.287	0.332	0.914*	0.845**	1.193***
Belgium	0.479**	0.376**	1.152***	1.145***	1.934***
Netherlands	-0.643***	-0.643***	0.055	0.050	0.626**
Germany	-0.187	-1.037***	0.005	0.012	0.416
Danemark	0.764***	0.764***	1.025***	1.469***	2.467***
Ireland	0.675	0.675	0.880*	1.276***	2.472***
United Kingdom	0.761**	0.281	0.338	0.888***	1.572***
Greece	2.667***	1.670***	2.302***		
Spain	1.446***	1.054***	1.640***		
Portugal	1.838***	0.958**	1.699***	1.562***	0.810*
Finland	2.812***	0.894***	1.420***	1.932***	2.094***
Sweden	1.600***	0.366*	1.053***	1.668***	2.439***
Austria	-0.638**	0.215	0.871**	1.145***	2.195***

*** p<0.01, ** p<0.05, * p<0.1. Italy is the omitted country.

Table 7. Description of some variables.

	Macroeconomic indicators		
Name	Description	Level	Source
Disposable income	Net disposable income of households expressed in PPP	NUTS II	Eurostat Regional statistics
Yearly change in disposable income	Change in net disposable income of households compared with the previous year	NUTS II	Eurostat Regional statistics
Population density	Inhabitants per km²	NUTS II	Eurostat Regional statistics
Unemployment	Unemployment rate	NUTS II	Eurostat Regional statistics
Suicides	Suicide rate (deaths per 100.000 inhabitants)	NUTS II	Eurostat Regional statistics
	Railways indicators		
Fares	revenue/passenger-km expressed in PPP	Country	World Bank's Railways database
Yearly change in fares	Change in fares compared with the previous year	Country	World Bank's Railways database
Accidents	Number of fatalities	Country	European Commission*
Monopoly	Number of providers of railway transport	Country	United Nations**
Demand	Passenger-km per capita	Country	World Bank's Railways database and Eurostat
Railways/motorways	Ratio between the length of railways and motorways	NUTS II	Eurostat Regional statistics

* Directorate-General for Energy and Transport
** Annual Bulletin of Transport Statistics for Europe and North America

Table 8. Some row data for the variables depicted in table 7.

	Population density*	Disposable income*	Unemployment rate*	Suicide rate*	Railways: monopoly**
Belgium	340.77	14.69	6.95	19.10	Yes
France	97.40	14.69	9.26	17.60	Yes
Netherlands	476.27	13.27	3.75	9.47	No
Germany	230.77	15.82	8.87	13.33	Yes
Italy	194.60	14.30	9.18	7.10	Yes
Denmark	124.63	11.33	4.65	12.87	Yes
Ireland	57.40	12.51	4.30	11.67	No
United Kingdom	243.57	15.69	5.06	7.07	No
Greece	84.07	11.46	10.40	3.20	No
Spain	81.87	12.38	12.06	8.27	Yes
Portugal	112.63	9.62	5.22	9.43	No
Finland	17.10	10.75	9.17	21.27	No
Sweden	21.77	12.34	5.86	12.90	Yes
Austria	98.07	15.78	4.14	18.70	No

* 2000, 2002 and 2004 average values
** 2004

Appendix 2.

Evidence presented in the first part of this work showed how perceived satisfaction differs between different groups of individuals. As we have seen, for example, elderly tend to be more satisfied than young people, women are less likely to be satisfied than men, center voters seem to appreciate rail transport more than others, etc.

The problem underlying this approach is that agents' satisfaction is based on both individual characteristics and on their travel experience. The former have been included in our models, whilst as a substitute for the latter we used some proxies as the level of fares, the ratio between railways and motorways, etc. This procedure allows to compare patterns of satisfaction among different countries but could lead to misleading results.

The evidence concerning the level of fares, for example, is somehow contradictory. As we mentioned in the previous section, a problem of endogeneity could affect our estimations, since the level of prices depends on a lot of unobserved factors which are supposedly to be included in our residuals. Moreover, each respondent based his/her judgement on his/her own travel experience, *i.e.* on the price he/she paid, on the delay he/she suffered, etc. An average measure or these factors obviously misses to capture such effects.

In order to avoid this problem we produce in this sub-section a new empirical application about consumers' satisfaction. Basically, we replicate the analysis carried out in the previous section by using the results from a survey conducted in June and July 2008. A sample of 407 travellers has been asked to answer a questionnaire about their satisfaction with rail transport in terms of punctuality. They have been also required to reveal some individual characteristics such as age, income and education.

The data.

Data originate from the survey about the estimation of the Value of Travel Time (VOT) presented in the previous chapter. Even if the first goal of this survey consisted in the analysis of travellers' preferences about time, we asked the respondents to state, on a scale from 1 to 10 their satisfaction with railway transport in terms of punctuality.
Respondents have also been asked about some individual characteristics such as age, gender, income, level of education and use of the transportation mode (journeys per month), reason of the journey (work or leisure trips). A more detailed description of the questionnaire is available at the end of this section, jointly with some descriptive statistics about the results.

The empirical model.

We analyse the connection between satisfaction and individual carachteristics from two different perspectives.

In the first one, assumptions about utility are the same as in the previous section. An ordered logit model is estimated on perceived quality. For the second sample the dependent variable has been recoded in four categories, as in Eurobarometer surveys. Those who assigned to punctuality a score between 1 and 3 are in the first category. The second one includes those who answered 4 or 5, while the third and the fourth categories comprehend respectively respondents who stated a judgement equal to 6 or 7 and equal or higher than 8.

In the second case we analyse the problem using a Bayesian Network (BN). Some theoretical issues about BN are summarised in the previous chapter (paragraph 3). Since the dataset is the same we used in the previous section, we estimate the same network, still using the *R* library *bnlearn* (Scutari, 2010).

Results.

Compared with the previous part of this work, here we are analysing a different group of individuals. In the first case, through Eurobarometer data, we were observing different patterns of satisfaction among respondents which depended on individual characteristics (partly observed) and on agents' travel experience (unobserved). Here, differences in perceived satisfaction still depends on individual characteristics, but this time all respondents shared the same travel experience, since they paid the same ticket, they suffered the same delay and they travelled on the same coach.

In the first case, even the effect of individual characteristics on the probability of being more or less satisfied could be biased by the fact that we do not have any information about respondents' travel experience. For example, recalling the results reported in table 6 in the previous section, we could suppose elderly to be less whining compared with young people. But it could also be the case that young individuals, supposed to be less wealthy than older respondents, usually travel on second class coaches, while elderly prefer first class. In this case different patterns of satisfaction across age could be simply explained by the fact that young and old respondents are judging two different things. The homogeneity of travel experience allows, in the second sample, to check the pure relation between individual characteristics and satisfaction.

Table 2 summarizes the results from the second sample. The first column shows the ordered logit estimates. The explanatory power of this model is very low. None of the individual characteristics is significant. A connection with satisfaction is found for other

two variables. In the survey, we asked the respondents if they have ever evaluated costs and benefits of other travel options (car, bus, etc.) before choosing the train. Those who answered positively, *i.e.* travellers who evaluated alternative options, tend to be less satisfied with the quality of the service in terms of punctuality. This relation could be explained by the assumption that those who carefully planned their trip and evaluated also other travel options pay more attention on delays.
Another question referred to the attitude of respondents to attach a monetary value to their time. Since this survey has been designed in order to estimate passengers VOT, it is quite surprising to note that a large share of respondents (about 33 per cent, table 1) do not consider at all delays as an economic damage. However, it is reasonable the sign of the coefficient reported in table 2: those who consider delays an economic damage tend to be more severe in evaluating service quality.

Since in the ordered logit model the parallel assumption is rejected, we estimated, as in the previous section, a generalised ordered logit, letting the coefficients of the troublesome variables free to vary across categories. The most significant effect is the one related to the gender, as females are more likely to be very unsatisfied with train punctuality than men. The relation between age and satisfaction is contradictory: more than on satisfaction, age of the respondents seems to have a link with the intensity of judgements.

Figure 1 reports the estimated BN. As we can see the network presents some differences compared with the ordered logit model. Satisfaction is directly affected by the reason of the trip (WORK). The parameter table (table 3) for the node SATISFACTION lists the conditional probability for the four outcomes of satisfaction given the reason of the trip. The difference between the two groups (work and non-work travellers) concentrates at the extreme: those who travel for non-work reasons are more likely to define themselves as “very satisfied” and less likely to report a judgement of strong dissatisfaction.

Even if the reason of the trip is the only variable which directly affects satisfaction, we cannot say the latter to be independent from other individual features such as age or gender. The structure of the network shows how the reason of the trip is influenced by the level of education which, in turn, is affected by both age and gender of the respondents. The same applies by the monetary value attached by each respondent to train delays (VALDELAY), since it is generated by the same family of arcs which lead to satisfaction

Conclusions.

The estimations focused on the second sample, in which all respondents experienced the same travel situation, does not provide strong evidence linking satisfaction and individual characteristics. This result could sounds like a contradiction compared with the findings analysed in the previous part of this analysis. However, we could ask

ourselves about the meaning of those findings. The evidence linking females to lower patterns of satisfaction is a strong result, confirmed by several contributions in the literature. The same applies for the result which links elderly to higher levels of satisfaction. However, what does this result mean? Why females should complain more than men?

The results presented in this appendix gives some hints about these questions and enable us to make a conjecture which needs further research in order to be confirmed. The ordered logit model and the Bayesian Network produced two results which share some interesting points in common. The main message they give us is that satisfaction is still affected by age and gender, but in an indirect way. Keeping constant the travel experience of respondents, they have an influence on satisfaction through attitudes, behaviours and ways of thinking, such as the reason of the trip, the way of evaluating travel alternatives or the attention they pay to delays and to other features of the service.

.

Table 1. Structure of the questionnaire and descriptive statistics.

		Absolute number	%
Gender (FEMALE)	Male	239	*58.72*
	Female	168	*41.28*
Reason for the trip (WORK)			
	Work trip	191	*46.93*
	Non work trip	216	*53.07*
How many times in a month do you travel on this rail line? (FREQ)			
	Less than one	177	*43.49*
	Between 1 and 3	158	*38.82*
	Between 3 and 5	41	*10.07*
	More than five	31	*7.62*
Age (AGE)			
	Less than 30	99	*24.32*
	Between 31 and 45	150	*36.86*
	Between 46 an 60	109	*26.78*
	Over 60	49	*12.04*
Education level (EDUC)			
	Elementary school	1	*0.25*
	Junior high school	27	*6.63*
	Senior high school	170	*41.77*
	University	209	*51.35*
Have you ever evaluated costs and benefits of other travel options (car, bus, etc.) for the trip between Turin and Milan? (ALTER)			
	Yes	177	*43.49*
	No	230	*56.51*
Why did you choose the railway? (REASON)			
	It is cheaper compared with other options	93	*22.85*
	It is faster compared with other options	17	*4.18*
	It is safer compared with other options	49	*12.04*
	I do not like to travel by car	52	*12.78*
	I do not have a car	71	*17.44*
	I like to travel by train / it is more comfortable compared with other options	125	*30.71*
Net yearly income (INC)			
	Less than 5.000	47	*11.55*
	Between 5.000 and 10.000	46	*11.30*
	Between 10.000 and 15.000	57	*14.00*
	Between 15.000 and 20.000	71	*17.44*
	Between 20.000 and 25.000	63	*15.48*
	Between 25.000 and 30.000	44	*10.81*
	Between 30.000 and 40.000	38	*9.34*
	Between 40.000 and 50.000	14	*3.44*
	Between 50.000 and 70.000	17	*4.18*
	More than 70.000	10	*2.46*

		Absolute number	%
How do you judge the quality of the service in terms of punctuality? (SATISFACTION)			
	1 (very bad)	4	*0.98*
	2	20	*4.91*
	3	33	*8.11*
	4	33	*8.11*
	5	47	*11.55*
	6	77	*18.92*
	7	84	*20.64*
	8	81	*19.90*
	9	16	*3.93*
	10 (excellent)	12	*2.95*
Do you consider train delays as an economic damage? (VALDELAY)	1 (severe damage)	2	*0.49*
	2	21	*5.16*
	3	43	*10.57*
	4	20	*4.91*
	5	42	*10.32*
	6	46	*11.30*
	7	27	*6.63*
	8	33	*8.11*
	9	39	*9.58*
	10 (not at all)	134	*32.92*
Have you ever tried to evaluate the damage in monetary terms?			
	Yes		
	No		

Table 2. Satisfaction with railway quality on the train between Turin and Milan. Ordered logit results (coefficients).

VARIABLES	Ordered logit	Generalized ordered logit		
		1	2	3
Age	0.006	-0.026**	-0.008	0.025***
Female	-0.313	-0.854***	-0.395*	-0.065
Education	-0.076	-0.071		
Income	0.006	0.015		
Work trip	-0.322	-0.335		
Frequency	-0.092	-0.096		
Evaluation of travel alternatives (yes)	-0.403**	-0.401**		
Reason of the travel choice				
Train is cheaper	-0.110	-0.063		
Train is faster	-0.099	-0.100		
Train is safer	0.471	0.455		
I do not like to travel by car	0.088	1.247**	0.250	-0.219
I do not have a car	0.063	0.019		
Economic value of delays	0.100***	0.101***		
Cut point (1)	-1.765**			
Cut point (2)	-0.567			
Cut point (3)	1.237			
Constant			3.233***	
Log-likelihood	*-517.164*		*-501.857*	
Chi-squared (Wald-chi squared)	*35.89*		*63.28*	
Observations	407		407	

Figure 1. Bayesian Network.

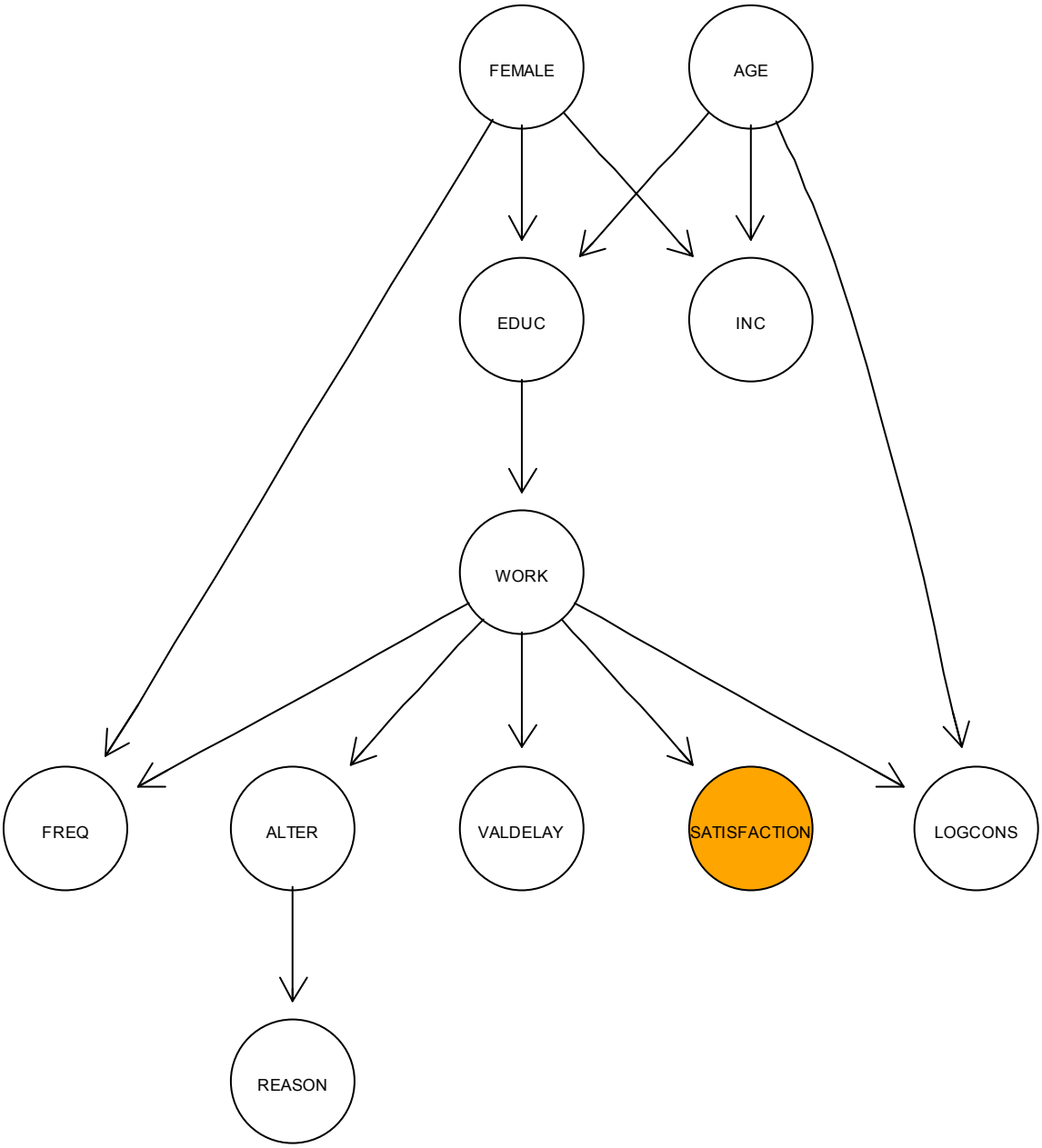


Table 3. Parameters of node SATISFACTION. Conditional probability table.

Satisfaction	Work trips	Non-work trips
Very satisfied	0.194	0.333
Fairly satisfied	0.408	0.384
Fairly unsatisfied	0.238	0.164
Very unsatisfied	0.162	0.120

References.

Bertrand M., Mullainathan S., 2001, *Do People Mean What They Say? Implications for Subjective Survey Data*, American Economic Review, 91(2), pp. 67-72.

Fiorio C., Florio M., 2008, *Do You Pay a Fair Price for Electricity? Consumers' Satisfaction and Utility Reform in the EU*, DEAS Working Papers 2008 – 12, Department of Economics, University of Milan.

Fiorio C., Florio M., Salini S., Ferrari P., 2007, *Consumers' Attitudes on services of General Interests in the EU: Accessibility, Price and Quality 2000-2004*, DEAS Working Papers 2007 – 2, Department of Economics, University of Milan.

FitzRoy F., Smith I., 1995, *The demand for rail transport in European countries*, Transport Policy, vol.2, n. 3, pp. 153-158.

Grassi S., Puglisi R., 2008, *Regulation and Consumers' Satisfaction from Public Services: an Individual Fixed Effect Approach*, DEAS Working Papers 2008 – 6, Department of Economics, University of Milan.

Hall J.A., Dornan M.C., (1990), *Patient sociodemographic characteristics as predictors of satisfaction with medical care: a meta analysis*, Social Science and Medicine, 30: 811-818.

Kahneman D., 2003, *Maps of Bounded Rationality: Psychology for Behavioral Economics, American Economic Review*, American Economic Association, vol. 93(5), pp. 1449-1475.

King G., Wand J., 2007, *Comparing incomparable survey responses: new tools for anchoring vignettes*, Political Analysis, 15, 1, pp. 46-66.

Koivumaa-Honkanen H., Honkanen R., Viinamaki H., Heiikila K., Kaprio J., Koskenvuo M., 2001, *Life satisfaction and suicide: a 20 year follow-up study*, The American Journal of Psychiatry, 158, pp. 433-439.

McFadden D., Bemmor A.C., Caro F.G., Dominitz J., Jun B., Lewbel A., Matzkin R.L., Molinari F., Schwarz N., Willis R.J., Winter J.K., 2005, *Statistical Analysis of Choice Experiments and Surveys*, Marketing Letters, 16 (3/4), pp. 183-196.

Rabin M., 2001, *A Perspective in Psychology and Economics*, UC Berkley Working Paper no. E02-313.

Scutari M., 2010, *Learning Bayesian Networks with the bnlearn R package*, Journal of Statistical Software, Vol. 35, n.3.

Sunstein C.H., Thaler R.H., 2003, *Libertarian Paternalism*, American Economic Review, n.2, pp.175-179.

Sunstein C.H., Thaler R.H., 2008, *Nudge – Improving Decisions About Health, Wealth and Happiness*, Yale University Press.

Williams R., 2006, *Generalized Ordered Logit/ Partial Proportional Odds Models for Ordinal Dependent Variables*, The Stata Journal 6(1):58-82.

Concluding remarks.

This work tried to analyse the issue of reliability in SP methods. Since SP studies and surveys are largely exploited in project evaluations and often use as guidelines for policy decisions, their reliability constitutes a relevant topic.

The first part of this work has been devoted to the discussion about the concept of reliability, pointing out how the economic literature deals with this issue and underlying some problematic aspects connected to the analysis of survey data.

We identified two main categories of survey. The first one includes WTP and WTA estimations, where people are required to simulate their own market behaviour in fictitious scenarios designed by the researcher. The second family involves those studies in which respondents are asked to reveal their attitudes and perceptions. The concept of reliability has been defined based on this distinction.

The second and the third section start from the definitions suggested in the first chapter and try to check SP reliability in two empirical studies.

The first one has been specifically carried out for this research. A sample of 407 train travellers has been asked to simulate their behaviour in several, hypothetical travel scenarios. The goal of this survey consisted in evaluating the WTP for travel time and both the structure of the questionnaire and the method used for the data collection reflect the ones usually used in this kind of analysis. Study's design allowed to define reliability in a less restrictive form compared with previous literature. Moreover, this formulation seems to fit better consumers' behaviour than the one applied in other works.

Results show how a significant share of respondents (about 25 per cent) was not able to properly report their preferences. WTP estimations demonstrate how the inclusion or the exclusion from the sample of unreliable choices produces significant implications for the exploitation of these data in policy evaluations.

A few contributions focused on this specific issue, also because the structure of the questionnaires usually allow for very restrictive logical test. As a consequence, the large share of inconsistent cases induces to doubt about the formulation of the test rather than about individuals' rationality. For this reason further research is needed, in order to fully understand how human limitations affect the correct estimations of WTP and WTA.

The second empirical work deals with two surveys of the second kind. Here respondents have not been asked to replicate their behaviour, but to express a judgement about their satisfaction with railway transport. The two surveys used in this section are somehow complementary. The first one is a collection of three Eurobarometer Issues. In this case

we have a large amount of information about respondents' characteristics, whilst we do not know many things about their travel behaviour. The second survey is, again, the study presented in the previous chapter. This time we know that all respondents shared the same travel experience (*i.e.* they paid the same fare, travelled on the same coaches, etc.), but we have less information about their individual features.

The joint analysis of these two datasets provides interesting results.

Through Eurbarometer data we saw how individual characteristics affect satisfaction even if we include in the model some control variables for objective quality. The evidence confirms the findings of previous literature on the subject. However, we could ask ourselves about the real meaning of these findings. Why females should be more satisfied than males? Why elderly should complaint less than young people? How gender and age affect satisfaction?

We have two possible explanations in order to answer these questions. The first one is to assume a sort of anthropological law which directly links individual characteristics with perceptions. The second, more realistic, one is to assume that individual characteristics affect perceptions indirectly, through attitudes, behaviours and ways of thinking which are not randomly distributed in the population but are, with a certain degree, group-specific.

Evidence form the second sample seems to confirm his conjecture. Age and gender still matter in defining patterns of satisfaction, but their influence affects perceptions through travel attitudes, such as the reason of the trip, the way of evaluating travel alternatives or the attention they pay to delays and to other features of the service.

Probably this conjecture merits further attention, since it would shed some light on the determinants of consumers' satisfaction and, consequently, on the reliability and proper interpretation of surveys' results.