Improving scope sensitivity in contingent valuation: joint and separate evaluation of health states

Pinto-Prades, José Luis ; Robles-Zurita, José Antonio; Sánchez-Martínez, Fernando-Ignacio ; Abellán-Perpiñán, José María ; Martínez-Pérez, Jorge

Published in:
Health Economics

DOI:
10.1002/hec.3508

Publication date:
2017

Document Version
Peer reviewed version

Improving scope sensitivity in Contingent Valuation: Joint and Separate Evaluation of Health States

José Luis Pinto-Prades, Glasgow Caledonian University, Cowcaddens Rd, Glasgow, Lanarkshire G4 0BA (United Kingdom), and University of Navarra, Campus Universitario. 31009 Pamplona. Navarra (España). JoseLuis.Pinto@gcu.ac.uk

José Antonio Robles-Zurita[1], Health Economics and Health Technology Assessment. Institute of Health and Wellbeing. University of Glasgow, 1 Lilybank Gardens, Glasgow, G12 8RZ, UK. JoseAntonio.Robles-Zurita@glasgow.ac.uk . Phone: +44 01413305615

Fernando-Ignacio Sánchez-Martínez, University of Murcia, Campus Universitario de Espinardo, s/n, 30100 Espinardo, Murcia (Spain). fernando@um.es

José María Abellán-Perpiñán, University of Murcia. dionisos@um.es

Jorge Martínez-Pérez, University of Murcia. jorgemp@um.es

Keywords: contingent valuation, evaluation mode, road safety, evaluability, health states.

Funding sources: Road Traffic Directorate General (*Dirección General de Tráfico,* unrestricted grant) and *Junta de Andalucía* (*proyecto de excelencia código* P09-SEJ-4992).

[1] Corresponding author.

Abstract

We present data of a contingent valuation (CV) survey, testing the effect of Evaluation Mode (EM) on the monetary valuation of preventing road accidents. Half of the interviewees was asked to state their Willingness to Pay (WTP) to reduce the risk of having only one type of injury (Separate Evaluation, SE), while the other half of the sample was asked to state their WTP for four types of injuries evaluated simultaneously (Joint Evaluation, JE). In the SE group we observed lack of sensitivity to scope while in the JE group WTP increased with the severity of the injury prevented. However, WTP values in this group were subject to context effects. Our results suggest that the traditional explanation of the disparity between SE and JE, namely, the so-called "Evaluability", does not apply here. The paper presents new explanations based on the role of preference imprecision.

Keywords: contingent valuation, evaluation mode, road safety, evaluability, health states.

1. INTRODUCTION

There is a debate about the validity of contingent valuation (CV) as an appropriate technique to inform social policies. While some critics (Hausman, 2012) think that it is a "hopeless" method, others (Carson, 2012) consider that, although the method is not perfect, it can be a useful technique to incorporate people's preferences in public decisions. An important part of the dispute focuses on the issue of scope effects. In order to improve the method, Heberlein et al. (2005) consider that "we need to better understand the conditions that produce scope failure" (p. 2). In this spirit, this paper focuses on the Evaluation Mode (Separate vs. Joint). We study whether evaluation mode makes a difference in the sensitivity of responses to scope in the specific domain of health state valuations.

There is a good deal of evidence (Hsee, 1996; Hsee et al., 1999; Hsee and Zhang, 2010; Bazerman et al., 1999) showing that subjects perceive he value of objects differently when they are presented in isolation (Separate Evaluation Mode –SE) or together (Joint Evaluation Mode -JE) and a mismatch between SE and JE valuations arise. More specifically, some individuals are willing to pay more for object A than for B when they are evaluated independently (SE) but are willing to pay more for B than for A when they are presented together (JE). This type of preference reversal has implications for the use of CV in public policy. Most public decisions involve choosing between alternative ways of spending a budget (i.e. Joint Evaluation Mode) while most CV studies elicit the monetary value of each policy independently from each other (i.e. Separate Evaluation Mode). If the values are different, which one (if any) should guide public policy?

The disparity between evaluation modes (EMs) has also been observed in the health domain (Lacey et al., 2006; Donaldson et al 2008; Gyrd-Hansen et al., 2011; Lacey et al., 2011) although only one of these papers (Donaldson et al 2008) deals with the monetary value of health. In Lacey et al. (2006) participants evaluated two health states, on a rating scale, using the two evaluation modes. They did not observe preference reversals but they found that the distance between the two health states was larger in JE than in SE. Gyrd-Hansen et al (2011) observed that subjects were more sensitive to the magnitude of risk reduction in JE than in SE. Thus both papers show that subjects are more sensitive to the magnitude of the object being evaluated in the JE mode. Donaldson et al (2008) estimated WTP for three different cancer programs (screening, treatment, rehabilitation) in different samples. Some subjects were asked to state their WTP for only one cancer program (SE) whereas some

other subjects were asked their WTP for two cancer programs (JE). They found that WTP changed with the EM and they attributed this result to the different amount of information that people have in each EM. Probably because of that explanation they seem to suggest that JE is a better EM when they stated that subjects in JE "will also understand better the respective impact of each of the programmes on their health" (p.5). We will offer in this paper a different explanation of the difference between EMs that does not lead so clearly to conclude that JE is a better EM. Moreover, the results of Donaldson et al (2008) do not shed light on the potential influence of EMs in the debate on scope effects since there was not any clear ranking between the three cancer programs. They were just different goods that did not differ on the amount of benefit provided (a priori). Some indirect evidence about the effect of the EM can be the literature on reference goods. Smith (2007) observed that subjects were willing to pay more for one health improvement when they were given information about the cost of an expensive intervention (the reference good) than when they were not given that information.

Given this evidence, we hypothesize that JE will increase sensitivity to scope in relation to SE. In this paper we present data of a large (n=2016) Computer Assisted Personal Interview (CAPI) survey aimed at obtaining the monetary value of the risk reduction of road traffic injuries of different severity. Half of the sample was asked to state their Willingness to Pay (WTP) to reduce the risk of having only one type of injury (SE group), while half of the sample was asked to state their WTP for four types of injuries evaluated simultaneously (JE group). The first contribution of this paper is providing evidence about the link between the EM and sensitivity to scope in a WTP study dealing with health outcomes. More specifically, we test the hypothesis that JE improves sensitivity to scope in relation to SE. The second contribution of the paper is providing a new theoretical interpretation of the reasons behind this result. We suggest that higher sensitivity to scope in JE can be due to the combined effect of preference imprecision and people's attempt to be internally consistent in their responses. This new theoretical interpretation is important because from showing that JE improves sensitivity to scope, it could be concluded that JE is a better EM. However, we will show that this conclusion is not so straightforward.

The paper is structured as follows. We first review the literature that relates EM and scope effects. Given that there is no evidence of this relationship in the health domain we will provide evidence gathered in other areas. This provides the theoretical framework of the

paper. Then we proceed to present the survey. In the fourth part we show the results. Although the main objective of the paper is to compare the two EMs in relation to scope effects, we also include an analysis of the results within JE, since we think this contributes toward a better understanding of the elements that influence responses. The discussion of results closes the paper.

2. EVALUATION MODES AND SCOPE EFFECTS IN CONTINGENT VALUATION

2.1. The effect of the Evaluation Mode

The literature about the effect of different EMs in CV studies is scant in economics. List (2002) asked subjects to state their monetary value of two different sets of baseball cards. One set of 10 cards (the "less" set) with a book value of about $15 and a set of 13 cards (the "more" set) comprising the same 10 cards as in the "less" set plus 3 additional cards of lower quality with a book value of $18. Subjects provided a higher monetary value to the "less" set than to the "more" set in SE but a lower monetary value in JE. This is the so-called "more is less phenomenon" (Hsee, 1998). This result was replicated in Alevy et al. (2011) and it was extended to environmental goods (wetlands clean-up and farmland preservation). In the case of wetlands the "less" group had to state their WTP for "an entire cleanup of 500 acres of wetlands" and in the "more" group the good to be valued was "an entire cleanup of 500 acres of wetlands and a partial cleanup of 50 acres". In the case of farmland the two goods were "permanently preserve 500 acres of Maryland farmland" and "permanently preserve 500 acres and temporarily (5 years) preserve 50 acres of Maryland farmland". Subjects were willing to pay the same for both goods in SE but they were willing to pay more for the good providing more benefit in JE. The effect in environmental goods was not as strong as with baseball cards, that is, instead of "more is less" they found that "more is the same". For this reason, Alevy et al. (2011) made a distinction between *strong EM effects* ("more is less") and *weak EM effect* ("more is the same"). Given that in both papers the results of JE are in line with normative theory (i.e. higher WTP for better goods) it could be thought that JE is a better EM. However, this depends on the way that those results are explained, as we show next.

The main explanation of the EM effect on preferences has been *Evaluability* (Hsee, 1996). In order to explain the concept of *Evaluability* and how it relates to scope effects we will

consider a model typically used in psychophysics and illustrated here in Figure 1.

INSERT FIGURE 1

Two functions are necessary to value an object using WTP (or any other response scale). One function (H) generates the impact of the object on the subjectivity of the individual (e.g. how well or badly this object is perceived). The other function (J) associates the response scale to the subjective impression. Hsee and Zhang (2010) define Evaluability as "the extent to which a person has relevant reference information to gauge the desirability of target values and map them onto evaluation" (pp. 344-345). This definition implies that Evaluability encompasses two different aspects: how easy it is for people to figure out how much utility an object is going to generate ("desirability") and how easy it is for people to translate ("map") this on the scale that is used to estimate the value of objects (money in CV). Desirability relates to the H function while mapping relates to the J function. We will show how these two elements of Evaluability relate to sensitivity to scope in JE. It is important to disentangle the origin of these effects since they may have implications for the normative status of each EM as a guide to public policy. One example of the use of JE vs. SE to disentangle the effect of the H and J functions in health is the study by Lacey et al (2011). They observed that patients and members of the general population value several health problems differently using a Visual Analogue Scale. They try to show if this disparity is produced by Visual Analogue Scale being used differently by the two groups (the J function) or because health is perceived differently (the H function).

2.2. Information Effects

The first reason that could lead to higher sensitivity to scope in JE is that in this EM subjects have more and better information to evaluate the quality of products. This helps subjects to understand more clearly how much utility an object can produce, how desirable it is (the H function) and how much they are willing to pay for the better object (scope effects). One reason that explains this effect is that some attributes are difficult to evaluate in isolation (in SE). One classic example (Hsee, 1996) is the choice between two dictionaries that are defined by two attributes, namely, the number of words and how new they look. The attribute that is easy to evaluate in SE is how new it looks while the number of words is difficult to evaluate in isolation. The consequence is that in SE the difficult-to-evaluate attribute is underweighted. However, in JE subjects can compare the number of

words of the dictionaries and it is easier for them to judge the quality of the dictionary by performing relative comparisons. In this case JE is more sensitive to scope (number of words) because it provides more "relevant reference information". This explanation is used by Lacey et al (2006) to explain some of their results when they state that the descriptions of health problems in JE provided "useful information about the range of severity that can be expected for the disease" (p.151). In the same way Gyrd-Hansen et al (2011) claim that the reduced sensitivity to differences in risk reduction in separate evaluations could be produced by the lack of comparators (i.e. lack of reference information). In the case of Smith (2007) this reference information is provided by the cost of the reference good. Donaldson et al (2008) conclude that "the main possibility of differences between JE and SE being due to informational effects" (p.15).

A second reason, also related to information, is that in JE subjects use wider frames in order to evaluate products. Assume that we evaluate two objects (A and B) and that A is, objectively, better than B. For example, A is a premium smartphone and B is a mid-range smartphone. However, assume that A is the worst within premium smartphones while B is the best within mid-range smartphones. Leclerc et al. (2005) show that in SE each object is evaluated within its category (what they call narrow focusing) leading to lower WTP for the best smartphone. This effect disappears in JE since subjects compare between mid-range and premium smart phones and are willing to pay more for the premium smartphone. That is, WTP reflects the objective ranking $A \succ B$. Again, if this is the explanation of the difference between SE and JE it seems logical to conclude that JE is a better EM to guide public policy. The disparity between EMs has also been explained in terms of a change in reference point (Leclerc et al., 2005). In SE each object is evaluated using its immediate category (e.g. premium smartphones) as the reference point. This implies that in SE each object is considered good or bad according to its ranking position in its own category. In JE each smartphone is compared against the other so the reference point is an object of a different category. This implies that subjects use a wider frame of reference in JE than in SE. It seems that this kind of argument is also used by Donaldson et al (2008) when they state that in SE subjects evaluate health programs in relation to *inappropriate reference points* while in JE a *relevant* alternative is presented. In summary, more information in JE leads to better reference points.

2.3. Imprecision/stochastic preferences

Differences between EM in CV studies may also reflect the difficulty that people have in measuring the desirability of an object with the money metric [J (X)]. Even if subjects have a good idea of how good an object is [H (X)] and attributes are evaluable in isolation, subjects may find it difficult to estimate with precision the monetary equivalent of the utility gain they can get from the consumption of some objects.

To explain how imprecision can account for discrepancies between both EMs, we assume that preferences are stochastic - the same subject might respond in a slightly different way to the same WTP question in different moments. We can think of individual preferences as a distribution of WTP values that the subject thinks are "reasonable" for an object (in our case to avoid a health problem). The WTP of one subject for object *g* will be defined as a random variable L^g, so $L^g = \{p_1^g, WTP_1^g; p_2^g, WTP_2^g; \ldots; p_n^g, WTP_n^g\}$ where $p_1^g \ldots p_n^g$ denotes the probabilities of stating a certain WTP amount (WTP_1, WTP_2, …, WTP_n) in a CV survey. We assume that the Expected Value $E[WTP(g)]$ of the distribution is the parameter that the CV survey has to estimate. We show next that if preferences are stochastic SE and JE can produce different results.

Assume that one subject responds to a WTP question for object *g* in SE mode. If her preferences are stochastic we assume that what the subject does is to choose one WTP value from L^g. Assume that, later on, she is asked a WTP question for object *f*. She responds choosing one number from L^f. Let us assume that (as will be the case in our study) *g* dominates *f*, that is, *g* is better than *f* in some dimensions and it is not worse than *f* in the rest of the dimensions (e.g. *f* is the "less" object and *g* is the "more" object). If there is some overlap between L^g and L^f then in SE, because of the degree of overlapping, $WTP^g < WTP^f$ could be observed. We hypothesise that the subject will not choose any pair (WTP^g, WTP^f) such that $WTP^g < WTP^f$ in JE since she will try to be internally consistent between the two WTP amounts stated. She may apply a social norm, in line with Norm Theory (Kahneman and Miller, 1986), that says you are expected to pay more for something that is better. If this is the case, subjects will not use the whole distributions L^g and L^f in JE when they respond to WTP questions. Subjects will truncate those distributions in order to avoid transparent violations of dominance (the social norm). The combined effect of stochastic preferences and the use of truncated distributions imply that the distance between $E[WTP(g)]$ and

$E[WTP(f)]$ will be larger in JE than in SE. Let us use an example to clarify this point. Assume that the probability distributions for *f* and *g* are, respectively, {4, 5, 6} and {5, 6, 7} with $p_1=p_2=p_3=1/3$ so $E[WTP(f)]=5$ and $E[WTP(g)]=6$ in SE. However, in JE subjects will only use WTP pairs that do not violate dominance. That is, [{4,5}, {4,6}, {4,7}, {5,6}, {5,7}, {6,7}]. This implies that $E[WTP(f)]=4.66$ and $E[WTP(g)]=6.3$ in JE. Furthermore, even if subjects are not sensitive to scope in SE and $L^g=L^f$, the theory just explained will predict that $E[WTP(g)]$ will be larger than $E[WTP(f)]$ in JE, indicating that we could observe sensitivity to scope in JE and insensitivity to scope in SE.

The idea that preferences are stochastic has a long tradition in economics (Mosteller and Nogee, 1951). Individual preferences are probabilistic and they are better represented by probability distributions than by a single value (deterministic preferences). There is evidence that moving from deterministic to stochastic preferences is all we need to explain some non-standard preferences. One example is Butler and Loomes (2007) who show how stochastic/imprecise preferences can explain preference reversals between matching and choice. Another example is Blavatskyy's (2007) truncated error model. This model explains violations of Expected Utility using two characteristics of preferences that we also use in this paper. One is that probability distributions can (sometimes) be truncated. The second one is that people do not commit transparent errors; for example people never choose a dominated alternative when dominance is transparent. Those assumptions can explain some biases in the way that people value objects. For example, assume that subjects have to state the monetary equivalent of a lottery with two monetary outcomes. Blavatskyy (2007) assumes that this monetary equivalent can be represented by a stochastic variable that is truncated by the two monetary outcomes of the lottery. Nobody will state a monetary equivalent larger than the highest outcome of the lottery or lower than the lowest outcome. This model implies that lotteries whose expected utility is close to the utility of the lowest possible outcome are more likely to be overvalued than undervalued (and vice versa). Similarly, our model assumes that imprecision and the attempt to be internally consistent leads to truncated distributions in JE as explained above.

In this section we have presented two reasons that can explain why JE can produce WTP values in line with sensitivity to scope; our study can also help to understand those reasons. If JE is more sensitive to scope because it provides the relevant information, the difference between SE and JE will vanish if we also give this information to those who are in SE. In

fact, there is some evidence that would support this explanation. Sher and McKenzie (2014) showed one group of subjects (group 1) objects A and B and they were asked to provide their WTP only for object A. They also presented objects A and B to another group (group 2) but they were asked their WTP only for object B. Finally, they asked another group (group 3) their WTP for objects A and B in JE. They found that WTP was the same in SE and in JE. This result is important since it suggests that giving more information led to more consistent results. Our second explanation in terms of stochastic preferences and internal consistency would not hold.

In summary, if the disparity between EMs disappears when subjects have the same information in SE and JE we can conclude that the difference between EMs is from varying information they convey. The implication would be that public policy should be based on WTP elicited in the JE mode or, at least, in SE mode subjects should be provided the same information received by those who are in JE mode. If the disparity between EM is not reduced when subjects have the same information in both EMs the implications are different. In this case, it is not so clear that JE is a better normative EM than SE. This paper aims at providing more evidence about the reasons of the relationship between the EMs and scope sensitivity that could serve as an input for a normative choice between EMs.

3. THE SURVEY

3.1. Participants and design

The survey was part of a project funded by the Spanish Ministry of Transport in order to estimate the value of non-fatal road injuries in road traffic accidents. A sample of 2016 subjects, representative of the Spanish adult general population were recruited. Respondents were selected by means of proportional stratified sampling by region, place of residence, sex and age of the respondent.

Eight different types of injuries (S1, S2, ..., S8) were selected based on Jones-Lee et al. (1995). Some minor modifications were made in order to produce dominance between all injuries. Dominance is interpreted here as a clear ranking in terms of severity, that is, $S1 \succcurlyeq \cdots \succcurlyeq S8$. The descriptions of the health states can be seen in the Appendix. These descriptions were presented to the respondents labelled as F, W, X, V, S, R, N and L, respectively, to avoid any suggested severity order.

The survey was administered through CAPI. The first part of the survey was an introduction that gave subjects information about the risk of road accidents in Spain. We also collected information about car use and attitudes toward road safety and perceptions about subjective risk.

Subjects were randomly allocated into 8 subgroups. Each group evaluated four of the eight different injuries using Ranking, Visual Analogue Scales and a Modified Standard Gamble (MSG) method before proceeding to the CV question(s) (see Table I). The rest of the questionnaire aimed at collecting socio-demographic information.

INSERT TABLE I

As shown in Table I, in all groups subjects had to rank four injuries as well as value them through the VAS and the MSG with the differences between SE and JE groups occurring in the CV tasks. In groups 1 to 4 (SE), respondents only saw the description of the injury they had to value using WTP. On the contrary, subjects in groups 5 to 8 (JE) were presented with the four health states they were going to value on the same screen, and then were asked their WTP to reduce the risk of each of the injuries.

3.2. Framing and CV elicitation

The Ranking task was very simple since subjects had to rank the health states from best to worst. Once they had ranked the four health states they had to value them on a line with the extremes identified as the “Best Imaginable Health State” (value 100) and the “Worst Imaginable Health State” (value 0). They also had to place "full health" and “death” on this scale and could say if some health states were so bad that they preferred to be dead rather than suffering those health states. After this task they had to evaluate the same four health states, randomly ordered, using a MSG. In this method, subjects are asked to choose between two lotteries. In one lottery, the outcomes are Full Health (FH) and Death (D), while in the other lottery, they are the health state to be evaluated (S1…S8) and Death (D). In the gamble with outcomes (S1…S8) and D the risk of death was fixed at 0.001 (1 in 1000), so lottery A is [0.999, S_i; D] i=1,…8. The probabilities (p) in the other lottery [p, FH; D] were adjusted until indifference was reached. Applications of the MSG are found in other studies (Carthy et al., 1998; Law et al., 1998; Bleichrodt et al., 2007; and Robinson et al., 2015). The relevant point for this paper is to stress that subjects were very familiar with the four health states they had to value in monetary terms before proceeding to the CV

questions, both in SE and in JE.

Figure 2 is a screenshot of the CV question for group 1 in SE. The task was explained to the subjects and they were only shown the description of the only injury they were going to value using WTP, in this case injury F (i.e. S1) (see left panel in Figure 2). They were told that there was a new safety device that could reduce injuries like F (in the example) in the case of a car accident from 15 to 10 in 100000. The safety device was personal and it had a lifespan of 1 year.

INSERT FIGURE 2

An example of the CV question is as follows[2]:

"Suppose you are offered a safety device, recently discovered, that can reduce the risk of health status F as a result of a traffic accident. This device, which is individual, can be used in any means of transport and has a lifespan of one year.

Suppose your risk of injury, such as F, as a result of a traffic accident is 15 in 100000 and that there exists a safety device that will reduce your risk of health status, such as F, in a traffic accident by 5 / 100000, from 15 in 100000 to 10 in 100000."

We used a set of payment cards in order to ask WTP questions. Each card represented an amount of Euros among these quantities: 10, 30, 50, 100, 150, 300, 600, 1000, 3000, 6000, 10000, 30000, 100000 and 300000. The method can be seen with the help of the right panel of Figure 2. A payment card showing a certain amount of money randomly appeared at the centre of the screen, and respondents had to assign the card to one of the next categories: a) "*I would pay this amount for sure"* (square at the right); b) "*I would not pay this amount for sure"* (square at the left) and; c) *"I am not sure whether I would pay or not"* (square at the bottom). For example, in Figure 2 a hypothetical respondent would definitely pay €50 or less and would definitely not pay €100 or more. When all the cards were allocated to the corresponding categories an open-ended question enquired about the maximum amount of money they would pay within the range defined by the highest amount

[2] In the introductory part of the survey a question was presented to subjects in order to check whether they understood risk ratios. The question was: *"Imagine that the probability of dying from a car accident is 1% (1 in 100 fatal accidents). In this situation, how many people would die for each group of 1,000?"* 97.17% of respondents answered the expected and correct answer (i.e. "10 people"). Then they were asked how many people would die for each group of 10,000. In this case 94,59% were correct (i.e. answered "100 people"). The huge majority, 94%, answered correctly both questions.

that they would pay for sure and the lowest amount that they would not pay for sure (in our example between €50 and €100). This open response is the WTP that we use in this study. During the whole process the description of the injury being valued was shown to the respondents on a paper card that was placed in front of them.

In JE subjects were first shown a screen with the four health states that they had to evaluate (Figure 3). It was explained that road traffic accidents could generate injuries of different severity and they were shown the four that they had already seen before in the VAS and in the MSG exercise. They were told that were going to be offered four different devices that could reduce the risk of having four different types of injuries. Each device could reduce the risk of one of those injuries. As in SE they were told that others could not use this device and the risk reduction was effective only over the next annual period. Then they moved to a sequence of four different screens. Each of the four screens was identical to the screen that was used to ask the WTP question in SE. The order of the injuries was random.

INSERT FIGURE 3

3.3. Hypotheses

This design makes it possible to test several hypotheses. If information is the explanation behind the disparity between SE and JE, we hypothesise that in our survey there will be no differences between EMs. That is,

H1: $WTP(S_i)_{SE}=WTP(S_i)_{JE}$ for i=1, 3, 4 and 6.

The reason for this hypothesis is that subjects in SE had the same relevant information as subjects in JE when they were asked the WTP question. All groups, in SE and JE had evaluated the same set of health states using different techniques (Ranking, Visual Analogue Scale and Standard Gamble) before the CV exercise so we assume that they had the same relevant reference information in both EMs.

If this hypothesis does not hold and WTP in SE and JE are different, the explanation in terms of Preference Imprecision and Internal Consistency can be tested. We then make the next hypotheses:

H2: $WTP(S1)_{JE}<WTP(S1)_{SE}$.

H3: $WTP(S3)_{JE} < WTP(S3)_{SE}$.

H4: $WTP(S6)_{JE} > WTP(S6)_{SE}$.

These come from the theory provided in section 2.3. Since S1 and S3 are the less severe health states in their respective groups in JE Preference Imprecision/Internal Consistency predicts that WTP distributions will be truncated from above (the part of the distribution with higher values). In the case of S6 it is the opposite. While for S4 no clear prediction can be made since it is in the middle and truncation can affect both sides of the distribution of WTP values.

4. RESULTS

4.1. Sample characteristics

Socio-demographic and attitudinal characteristics of our sample can be seen in Table II for the total sample and for each of the eight groups. We also show the distribution of adult population with respect to age and gender, according to the Spanish 2011 census, and with respect to education, marital status and employment status, according to the Labour Force Survey (LFS).[3] In general, our sample resembles the characteristics of the population. More information was collected about other characteristics as shown in Table II. We performed a Chi^2 test for independence between groups and each of the characteristics. We could only reject the null hypothesis for employment status at 5% of error. All the remaining characteristics appear to be equally distributed among groups.

INSERT TABLE II

4.2. Testing the hypotheses

The impact of the EM on WTP can be seen in Table . We deal with outliers in two different ways. The first one is trimming, specifically we trimmed the top 2% of the values (5 observations per group). The second is winsorization (Kahneman and Ritov, 1994), that is the 12 highest observations (about 5% of each group) were substituted with the value of the 13th highest one. On the lower part of the scale nothing was changed since the 13th lowest observation always coincided with the 12 previous observations (they were 0). We prefer to

[3] See report on the 1st quarter of the 2011 Spanish Labor Force Survey in: http://www.ine.es/daco/daco42/daco4211/epa0111.pdf.

present the results using winsorization because it does not change the shape of the distribution. Nonetheless the results of the statistical tests are the same using winsorization or removing 5 outliers. Means and medians are also very similar with the two strategies we used to deal with outliers.

INSERT TABLE III

We can see that means and medians follow the expected pattern (the higher the severity of the health state, the higher the WTP) in JE. In SE medians are the same for S1 and S3 and they are also the same for S4 and S6. This suggests insensitivity to scope in SE for some comparisons. In SE there were no statistically significant differences between S1 and S3 or between S4 and S6 showing lack of scope sensitivity. However, statistically significant differences (p-value<0.01) were found between all health states in JE.

It can be seen that H1 does not hold in most of the cases. Only in two cases -S3 and S4 (group 7)- there are no differences between SE and JE. In the rest of the cases, differences are in line with hypotheses H2, H3 and H4. H2 holds in the two tests, that is, $S1_{SE}$is greater than $S1_{JE}^{G5}$ and $S1_{JE}^{G8}$. H3 holds when $S3_{SE}$ is compared with $S3_{JE}^{G6}$. H4 holds in the two tests. However, in the case of S4 we do not observe a clear pattern, as we could expect from theory. In one case the WTP is higher in SE than in JE and in the other case it is the opposite. Overall, the results suggest that the explanation of the data in terms of Imprecision plus Internal Consistency seems plausible. Information cannot account for differences between EMs given that it was the same in all groups.

In relation to sensitivity to scope, we can see that, in SE, subjects were not sensitive to scope between health states S1 and S3 while in JE, they were sensitive to scope in all cases. However, even when subjects were sensitive to scope in SE there were clear differences in the values elicited with SE or JE. There is much more overlap in the WTP distributions in SE than in JE. This is clearly seen in the box-plot diagram in Figure 4. The consequence of the above is that the relative values of health states are very different in the two EMs. In JE the ratios of WTP values are more extreme. For example, in JE the ratio of means between injury S1 and S6 is 8.11 (€635.9/€78.4) while in SE is 2.4 (€436.7/€181.3). This implies that avoiding one injury like S6 is equivalent to avoiding 8 injuries like S1 using JE but only 2.4 injuries like S1 using SE.

INSERT FIGURE 4

4.3. Further results

Other results suggest that WTP values in JE are influenced to some extent by some kind of strategy used by subjects to be internally consistent. We can see (Table IV) that in almost all cases the differences between WTP values are statistically significant *within* each group even if health states are not too different (e.g. S1 and S2). However, there were several cases where the differences did not reach statistical significance when health states were compared between groups even within JE.

INSERT TABLE IV

Another result that adds to this evidence is presented in Table V. We show the percentage of subjects who made a mistake (reported a higher WTP for the less severe health state) and the percentage of subjects who reported exactly the same WTP. Those results suggest some kind of process to be internally consistent. If subjects had responded to each WTP question independently from each other, we would have observed a fair amount of errors for similar health states (e.g. S1 vs. S2) and almost no errors for very different health states (e.g. S1 vs. S8). Errors should have been inversely related to the difference between the severity of health states. We do not observe anything like that. Instead, we see almost no errors in all cases, no matter how similar or dissimilar the health states are, and a large number of subjects providing exactly the same response for health states that are different. We interpret that result as evidence of Internal Consistency. That is, subjects are not sure about which is their true WTP but they understand that it is illogical to pay more for something worse (the Norm). However it appears that they do not see anything wrong in providing the same response to two different health states.

INSERT TABLE V

4. DISCUSSION

We have seen that the values elicited for different health states change with the EM used to elicit preferences. More specifically, we have seen that in SE subjects are (to some extent) insensitive to scope. We have also seen that in JE subjects discriminate more between health states. Similar results have also been observed in Lacey et al (2006) and Gyrd-

Hansen et al (2011). Lacey et al (2006) find that the difference between the value of a mild and a severe lung problem increased from 21 points in SE to 54 points in JE on a 0-100 rating scale. Gyrd-Hansen et al (2011) observed that the differences between two different risk reductions were higher in JE than in SE. Another result, in line with our findings, in Gyrd-Hansen et al (2011) is that 52.5% of subjects in JE gave the same value to two different risk reductions but nobody gave a higher value to the smallest risk reduction. This suggests some kind of effort from subjects to be internally consistent, as it also seems to be happening in our study.

Differences between EMs have usually been attributed to informational effects (*Evaluability*). Donaldson et al. (2008) find that WTP for a cancer screening program is more likely to be higher when elicited together with a treatment or a rehabilitation programme (JE). Similar results are obtained, though less conclusive, for the treatment and rehabilitation programmes and they attribute these findings to informational effects. We argue in this paper that *Evaluability* does not seem to be the only explanation of the disparity between EMs. In fact, in the case of health states we could assume that subjects should be more or less familiar with the severity of health outcomes. As Lacey et al (2006) say "in the case of our lung disease scenarios, the evaluability of lung disease severity should not have been especially poor" (p. 151). We think that most subjects would be able to think of a mild headache as a mild health problem and of a metastatic cancer as a very severe problem without the need of the information provided by the study. This is why it is important to explain the effect of Response Mode in a different way, as done in this paper. We present a complementary explanation based on the stochastic nature of human preferences combined with the attempt to be internally consistent.

We started our paper asking if CV can be improved using JE Mode. If by "improving" we mean to produce values that are more sensitive to scope, the answer is affirmative: JE produces values that are more in line with what we would expect from theory. However, this cannot be attributed to improved *Evaluability*, i.e. subjects understanding better how severe a health problem is. Part of the explanation of the scope effects that we have observed in JE seems to reflect the adjustments that subjects make in order to be internally consistent. What are the implications of this finding for CV? Should we elicit WTP values in SE or in JE?

We think that there are several ways to respond to those questions depending on views

about preferences. Under the assumption that social policy should be based on consistent and stable individual preferences, our results could be read as supporting the idea that CV cannot be trusted. In theory, preferences should not depend on the EM. However, we see that they do vary depending on the EM. A different approach is to accept that WTP questions for risk reductions are difficult for members of the general population and that, even if they are imprecise and stochastic, they provide valuable information for the social decision maker. For those who hold this second view, we provide some arguments to defend the use of JE.

First, we may think that the internal consistency observed in JE is an example of Coherent Arbitrariness (Ariely et al., 2003) and it does not provide any evidence for the superiority of JE over SE. However, we do not think that WTP responses in JE can necessarily be understood as "arbitrary". In Experiment 1 of Ariely et al. (2003) the first response is considered "arbitrary" because subjects are influenced by a random (arbitrary) number (the last two digits of their Social Security number). This does not have to be the case in our study. We can assume that the first response comes from a set of values that are all "reasonable" (or "true") for the subject. Their response is stochastic but not arbitrary. Also, the social norm that regulates the second response is not arbitrary but normatively appealing. The fact that subjects try to be internally consistent in JE does not imply that the values elicited using JE are totally arbitrary or that they do not have normative status. This is the view of Frederick and Fischhoff (1998) when they write, "*we do not believe that the demands of within-subject designs necessarily decrease the validity of the contingent responses – indeed, respondents in a within–subject quantity manipulation who report that a lot more of a good is worth a lot more to them may be revealing more about their true values than respondents in a between-subject design, who (collectively) indicate that a lot more of a good is only worth only a little more*" (p. 116). More recently, Kahneman (2014) presented some reasons that also support the use of JE to guide social policy (and even individual decisions). He uses the example (taken from Johnson et al., 1993) of a study where subjects were asked in SE (between-subjects) their WTP for two insurance policies. In one group, the insurance policy paid $100,000 if the subject died, for any reason, during a holiday trip. In the second group, they were asked the WTP question for a similar policy that only paid $100,000 if the subject died, due to a terrorist attack, during the holiday trip. Subjects were willing to pay more for the second insurance policy. Kahneman (2014) argues that if people had been able to compare the two policies (that is, if they had evaluated the two policies in JE) they would

have seen that the first gave a higher value to them and they would have been willing to pay more for the first policy. He attributes this disparity between EMs to what he calls "substitution". In SE subjects respond to a different question to the one they are being asked (substitution). Since subjects find it very difficult to imagine how much they are willing to pay for an insurance policy they respond emotionally; that is, they are willing to pay more for the second policy because they are more afraid of dying in a terrorist attack than of death itself. However, in JE they realise that it does not make sense to pay more for a policy that offers less protection. That is, under JE they would taken the best decision.

We can use this analogy when we move to health problems (this paper). When subjects are asked their WTP to reduce their risk of a certain injury they respond according to the degree of fear that the injury generates. This argument that the valuation of health states mainly reflects the degree of "shock reaction to, or fear associated with, that state" (p. 223) has been used by Dolan and Kahneman (2008, p. 223). It could be the case that states S1 and S3 generate the same degree of (low) fear since they are both mild. In the same way, more severe health states generate more fear and this increases WTP. However, there are reasons to think that preferences based on the intensity of emotions do not seem to be a good guide for social policy (Slovic et al., 2004 and 2005; Finucane et al., 2000) because feelings are very often not well correlated with benefits. We can think that JE requires subjects to think more rationally and less emotionally and this corrects the lack of scope effect that we can see in SE; the fact that subjects adjust their responses in order to be internally consistent in JE is also a manifestation of their preferences. Subjects realise that it does not make sense to pay more for avoiding a health state that is less severe than another one. Subjects may not know very well what is the right WTP for S1 or S3 (stochastic preferences) but they know that WTP for S1 cannot be bigger than for S3. In that respect, as Frederick and Fischhoff (1998) state, in JE subjects "may be revealing more about their true values" than in SE. In fact, the results obtained in JE seem to have better properties to guide social policy than the results obtained in SE. It does not make sense to accept that subjects' true preferences are that S1 and S3 are equally bad (as SE evaluations suggest) or that the benefit of preventing 10 injuries like S6 is equivalent only to preventing 24 like S1. At face value, this seems implausible, given how different they are. We conclude that, in the presence of imprecise/stochastic preferences, JE can be a better EM than SE and that CV can be improved using this response mode.

Acknowledgements

The authors want to express their gratitude to participants at the 34th Spanish Health Economics Conference, Pamplona, and to assistants to a seminar at Yunus Centre for Social Business and Health, Glasgow Caledonian University. The authors thank the Spanish Road Traffic Directorate General (*Dirección General de Tráfico*), for an unrestricted grant, and *Junta de Andalucía* (*proyecto de excelencia código* P09-SEJ-4992).

REFERENCES

Alevy, J. E., List, J. A., and Adamowicz, W. L. (2011). How Can Behavioral Economics Inform Nonmarket Valuation? An Example from the Preference Reversal Literature. *Land Economics*, *87*(3), 365-381.

Ariely, D., Loewenstein, G., and Prelec D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. The *Quarterly Journal of Economics*, 118: 73-105.

Bazerman, M. H., Moore, D. A., Tenbrunsel, A. E., Wade-Benzoni, K. A., and Blount, S. (1999). Explaining how preferences change across joint versus separate evaluation. *Journal of Economic Behavior and Organization*, *39*(1), 41-58.

Blavatskyy, P. R. (2007). Stochastic expected utility theory. *Journal of Risk and Uncertainty*, *34*(3), 259-286.

Bleichrodt, H., Abellan-Perpiñan, J. M., Pinto-Prades, J. L., and Mendez-Martinez, I. (2007).Resolving inconsistencies in utility measurement under risk: tests of generalizations of expected utility. *Management Science*, *53*(3), 469-482.

Butler, D. J and Loomes, G. C. 2007. Imprecision as an account of the preference reversal phenomenon. *American Economic Review* 97 (1): 277-297

Carson, R. T. (2012). Contingent valuation: A practical alternative when prices aren't available. *The Journal of Economic Perspectives: A Journal of the American Economic Association*, *26*(4), 27-42.

Carthy, T., Chilton, S., Covey, J., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon, N.,and Spencer, A. (1998). On the contingent valuation of safety and the safety of contingent valuation: Part 2-The CV/SG" chained" approach. *Journal of risk and uncertainty*, *17*(3), 187-214.

Dolan, P., and Kahneman, D. (2008).Interpretations of utility and their implications for the valuation of health. *The Economic Journal*, *118*(525), 215-234.

Donaldson, C., Luchini, S., Moatti, J.P., and Protière, C. (2008). Contingent Valuation of Competing Public Sector Programmes: An Experiment of Single vesus Joint Evaluation. Document de Travail nº 2008-27. GREQAM. [available at: https://halshs.archives-ouvertes.fr/halshs-00282844/document]

Finucane, M. L., Alhakami, A., Slovic, P., and Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits.*Journal of Behavioral Decision Making*, 13, 1-17.

Frederick, S., and Fischhoff, B. (1998).Scope (in) sensitivity in elicited valuations. *Risk Decision and Policy*, *3*(2), 109-123.

Gyrd-Hansen, D., Halvorsen, P., Nexøe, J., Nielsen, J., Støvring, H., and Kristiansen, I. (2011). Joint and Separate Evaluation of Risk Reduction Impact on Sensitivity to Risk Reduction Magnitude in the Context of 4 Different Risk Information Formats. *Medical Decision Making*, *31*(1), E1-E10.

Hausman, J. (2012). Contingent valuation: From dubious to hopeless. *The Journal of Economic Perspectives: A Journal of the American Economic Association*, *26*(4), 43-56.

Heberlein, T. A., Wilson, M. A., Bishop, R. C., and Schaeffer, N. C. (2005). Rethinking the scope test as a criterion for validity in contingent valuation. *Journal of Environmental Economics and Management*, *50*(1), 1-22.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, *67*(3), 247-257.

Hsee, C. K. (1998). Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, *11*(2), 107-121.

Hsee, C. K., Loewenstein, G. F., Blount, S., and Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576.

Hsee, C. K., and Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science*, *5*(4), 343.

*Instituto Nacional de Estadística (2011). Encuesta de Población Activa. Primer trimestre de 2011.*http://www.ine.es/daco/daco42/daco4211/epa0111.pdf (last access 21-07-2015).

Johnson, Eric J, Hershey,J. Meszaros, J. and Kunreuther, H. (1993) Framing, Probability Distortions, and Insurance Decisions. *Journal of Risk and Uncertainty*, 7: 35-31.

Jones-Lee, M. W., Loomes, G., and Philips, P. R. (1995).Valuing the prevention of non-fatal road injuries: Contingent valuation vs. standard gambles. *Oxford Economic Papers*, 676-695.

Kahneman D. (2014) Hindsights, Public Lecture Series, Princeton (10th February).

Kahneman D., and Miller D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, Vol 93(2): 136-153.

Kahneman, D., and Ritov, I. (1994).Determinants of stated willingness to pay for public goods: A study in the headline method. Journal of Risk and Uncertainty, 9(1), 5-37.

Lacey, H. P., Fagerlin, A., Loewenstein, G., Smith, D. M., Riis, J., andUbel, P. A. (2006). It must be awful for them: perspective and task context affects ratings for health conditions. *Judgment and Decision making*, *1*(2), 146-152.

Lacey, H. P., Loewenstein, G., andUbel, P. A. (2011).Compared to what? A joint evaluation method for assessing quality of life. *Quality of Life Research*, *20*(8), 1169-1177.

Law, A. V., Pathak, D. S., and McCord, M. R. (1998). Health status utility assessment by standard gamble: A comparison of the probability equivalence and the lottery equivalence approaches. *Pharmaceutical research*, *15*(1), 105-109.

Leclerc, F., Hsee, C. K., & Nunes, J. C. (2005).Narrow focusing: Why the relative position of a good in its category matters more than it should. *Marketing Science*, *24*(2), 194-205.

List, J. A. (2002). Preference reversals of a different kind: The "more is less" phenomenon. *American Economic Review*, *92*(5), 1636-1643.

Mosteller, G., Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy* 59 (5): 371-404

Robinson, A., Spencer, A., and Moffatt, P. (2015). A framework for estimating health state utility values within a discrete choice experiment modelling risky choices. *Medical Decision Making*, *35*(3), 341-350.

Sher, S., and McKenzie, C. R. (2014).Options as information: Rational reversals of evaluation and preference. *Journal of Experimental Psychology. General*, *143*(3), 1127-43.

Slovic, P., Peters, E., Finucane, M. L., and Macgregor, D. G. (2005). Affect, risk, and decision making. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *24*(4Suppl), S35-40.

Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, *24*(2), 311-22.

Smith, R. (2007). The role of ‘reference goods’ in contingent valuation: should we help respondents to ‘construct’ their willingness to pay? *Health Economics*, 16(12), 1319-1332.

Figures and tables

Figure 1. Evaluation model

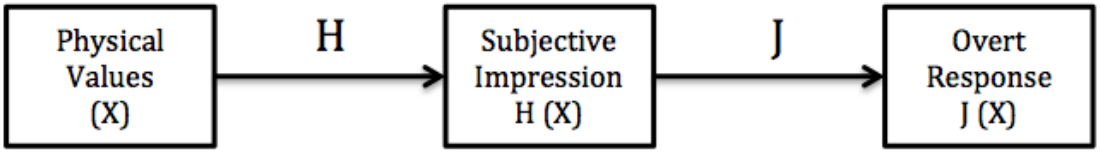
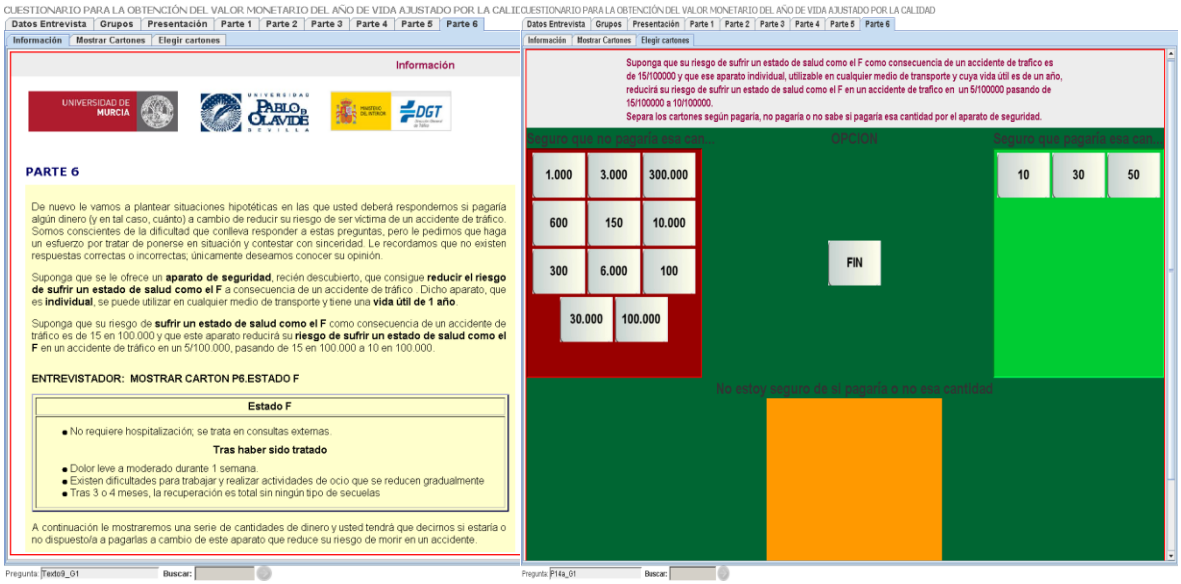
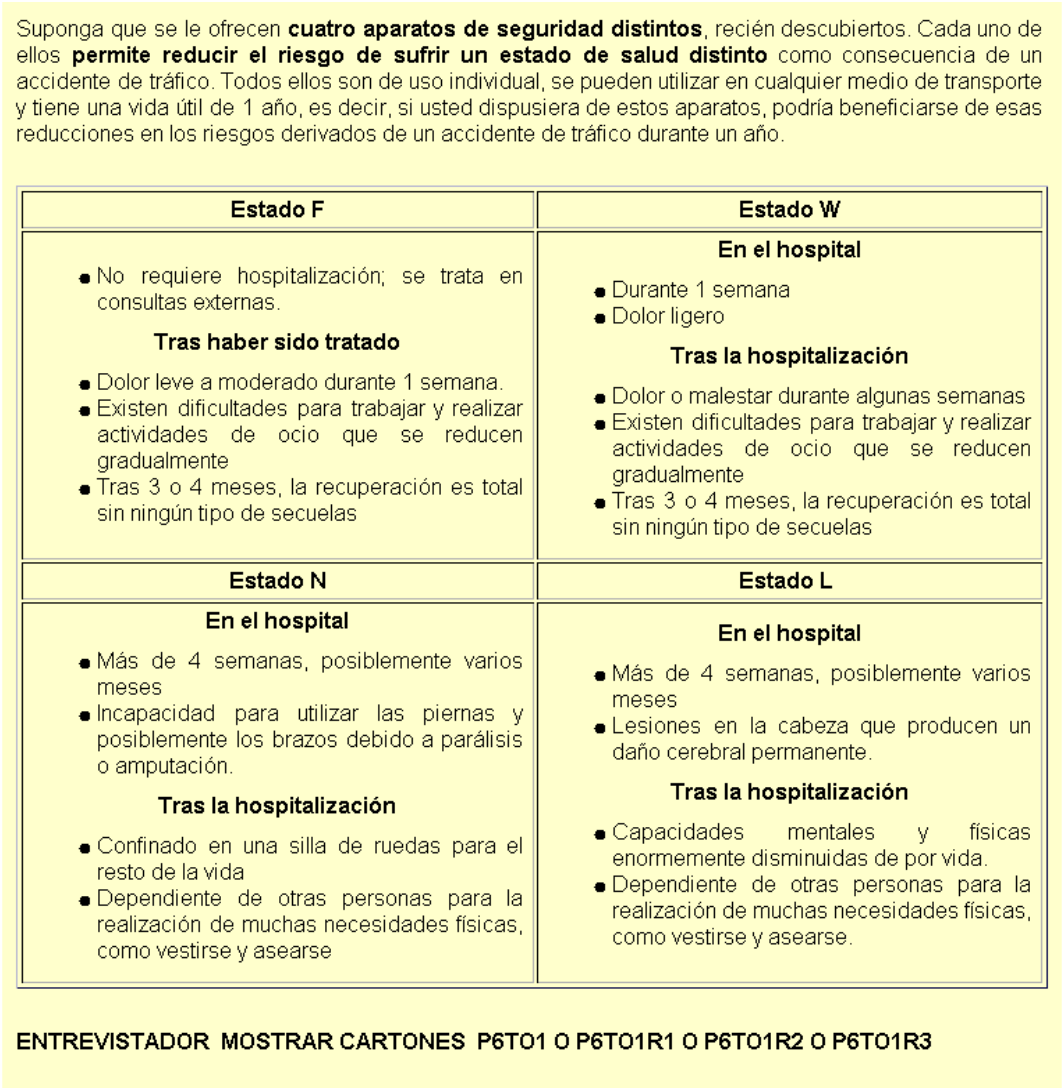


Figure 2. Screenshot for CV question in SE, Group 1



Note. Find description of health state F in English in the appendix.

Figure 3. Screenshot for CV question in JE, Group 5



Suponga que se le ofrecen **cuatro aparatos de seguridad distintos**, recién descubiertos. Cada uno de ellos **permite reducir el riesgo de sufrir un estado de salud distinto** como consecuencia de un accidente de tráfico. Todos ellos son de uso individual, se pueden utilizar en cualquier medio de transporte y tiene una vida útil de 1 año, es decir, si usted dispusiera de estos aparatos, podría beneficiarse de esas reducciones en los riesgos derivados de un accidente de tráfico durante un año.

Estado F	Estado W
• No requiere hospitalización; se trata en consultas externas. **Tras haber sido tratado** • Dolor leve a moderado durante 1 semana. • Existen dificultades para trabajar y realizar actividades de ocio que se reducen gradualmente • Tras 3 o 4 meses, la recuperación es total sin ningún tipo de secuelas	**En el hospital** • Durante 1 semana • Dolor ligero **Tras la hospitalización** • Dolor o malestar durante algunas semanas • Existen dificultades para trabajar y realizar actividades de ocio que se reducen gradualmente • Tras 3 o 4 meses, la recuperación es total sin ningún tipo de secuelas
Estado N	**Estado L**
En el hospital • Más de 4 semanas, posiblemente varios meses • Incapacidad para utilizar las piernas y posiblemente los brazos debido a parálisis o amputación. **Tras la hospitalización** • Confinado en una silla de ruedas para el resto de la vida • Dependiente de otras personas para la realización de muchas necesidades físicas, como vestirse y asearse	**En el hospital** • Más de 4 semanas, posiblemente varios meses • Lesiones en la cabeza que producen un daño cerebral permanente. **Tras la hospitalización** • Capacidades mentales y físicas enormemente disminuidas de por vida. • Dependiente de otras personas para la realización de muchas necesidades físicas, como vestirse y asearse.

ENTREVISTADOR MOSTRAR CARTONES P6TO1 O P6TO1R1 O P6TO1R2 O P6TO1R3

Note. Find description of health states in English in the appendix.

Figure 4.Box-Plot of WTP in each Evaluation Mode.

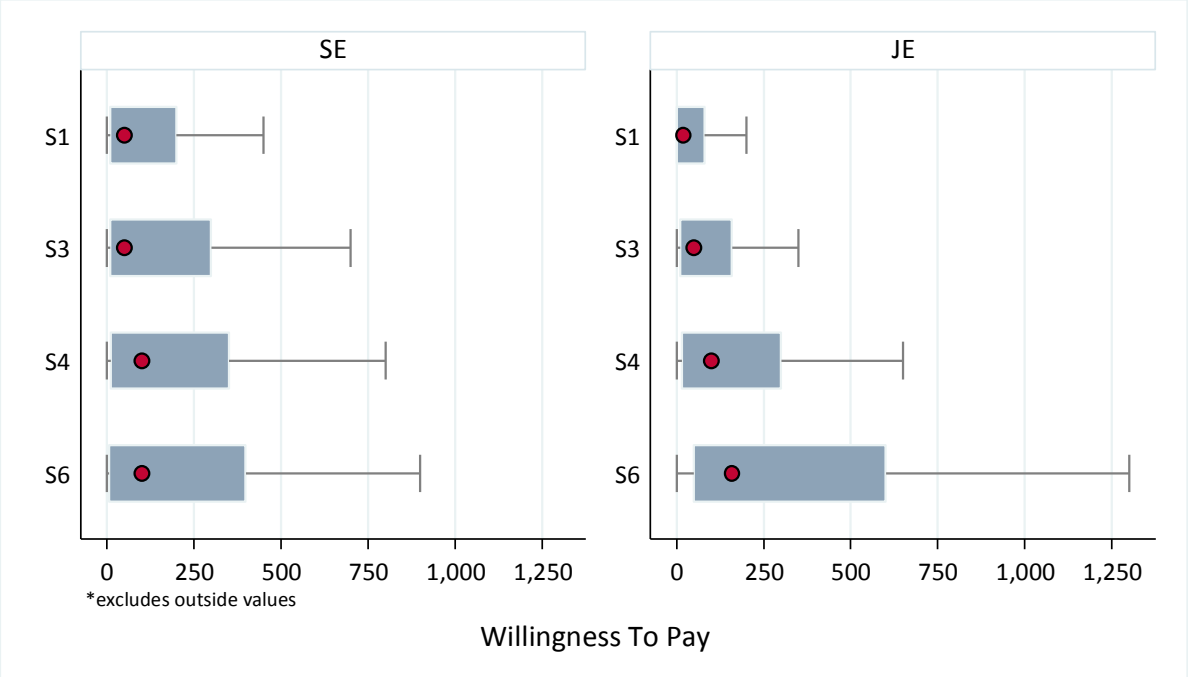


Table I. Survey Design

Evaluation Mode	Group	N	Health states in Ranking, Visual Analogue Scale and MSG	Health states in CV
SE	1	254	S1, S2, S7, S8	S1
	2	251	S3, S4, S7, S8	S4
	3	256	S3, S4, S5, S6	S3
	4	251	S1, S2, S5, S6	S6
JE	5	253	S1, S2, S7, S8	S1, S2, S7, S8
	6	250	S3, S4, S7, S8	S3, S4, S7, S8
	7	248	S3, S4, S5, S6	S3, S4, S5, S6
	8	253	S1, S2, S5, S6	S1, S2, S5, S6

Table II. Percentage distribution of sample characteristics

Variables	Total sample	By group 1	2	3	4	5	6	7	8	Chi^2 test (p-value)	Population
Gender											**Census**[a]
Male	48.8	49.6	47.4	48.1	49.0	48.2	50.0	49.2	48.6	0.999	49.3
Female	51.2	50.4	52.6	52.0	51.0	51.8	50.0	50.8	51.4		50.6
Age											**Census**
18-29	17.8	18.9	18.7	18.4	18.3	14.2	19.2	16.9	17.4	0.999	16.1
30-39	20.3	19.3	20.7	18.8	20.7	23.3	18.8	23.0	18.2		20.2
40-49	20.7	20.9	21.1	19.5	21.5	20.2	19.2	19.4	24.1		19.4
50-65	23.7	22.1	23.1	23.8	22.3	25.3	25.6	24.2	23.3		23.3
>=66	17.5	18.9	16.3	19.5	17.1	17.0	17.2	16.5	17.0		20.9
Education											**LFS**[b]
No ed., Prim. or Lower Sec.	50.8	52.8	45.8	46.1	53.8	49.8	54.0	53.2	51.4	0.736	54.8
Upper Secondary	25.8	22.1	28.3	28.1	25.1	25.7	25.6	25.8	25.7		20.3
Tertiary	23.4	25.2	25.9	25.8	21.1	24.5	20.4	21.0	22.9		24.8
Employment Status											**LFS**
Inactive	40.1	37.8	35.9	46.1	46.6	39.1	42.0	32.7	40.3	0.034	40.1
Employed	47.8	49.6	49.4	44.9	45.4	47.4	42.8	53.2	49.4		47.1
Unemployed	12.2	12.6	14.7	9.0	8.0	13.4	15.2	14.1	10.3		12.7
Marital status											**LFS**
Single	23.8	27.6	23.1	25.8	25.5	21.3	18.0	23.8	24.9	0.475	31.4
Married	63.6	60.6	62.2	60.2	64.1	65.6	71.6	62.9	62.1		56.5
Divorced	5.8	3.9	7.6	5.5	4.8	7.5	5.2	6.9	4.7		4.8
Widow	6.9	7.9	7.2	8.6	5.6	5.5	5.2	6.5	8.3		7.3
Household income (€)											
0 – 1,200	49.3	48.0	46.6	51.2	46.2	52.6	50.8	50.0	48.6	0.318	
1,201 – 1,800	25.3	23.2	22.3	23.8	25.5	28.5	27.6	24.6	26.9		
>1,800	25.5	28.7	31.1	25.0	28.3	19.0	21.6	25.4	24.5		
Smoker											
Non smoker	66.8	65.4	64.5	70.3	64.1	64.8	71.6	63.7	69.6	0.331	
Smoker	33.2	34.7	35.5	29.7	35.9	35.2	28.4	36.3	30.4		
Alcohol											
No	41.6	34.7	40.6	43.0	47.0	41.5	44.4	39.1	42.3	0.196	
Yes	58.4	65.4	59.4	57.0	53.0	58.5	55.6	60.9	57.7		

a. Spanish 2011 census.
b. 1st quarter of Spanish 2011 Labour Force Survey.

Table II. Percentage distribution of sample characteristics (Cont.)

Variables	Total sample	By group 1	2	3	4	5	6	7	8	Chi^2 test (p-value)	Population
Practices sports											
No	43.3	41.7	39.8	40.2	44.2	50.2	44.0	40.7	45.5	0.279	
Yes	56.7	58.3	60.2	59.8	55.8	49.8	56.0	59.3	54.6		
Driver											
No	38.2	35.0	35.5	39.1	39.0	38.3	44.0	33.9	41.1	0.293	
Yes	61.8	65.0	64.5	60.9	61.0	61.7	56.0	66.1	58.9		
Gambles											
No	29.6	26.8	27.9	30.1	28.7	32.8	34.4	24.6	31.6	0.370	
Yes	70.4	73.2	72.1	69.9	71.3	67.2	65.6	75.4	68.4		
Self-reported Health											
Excellent	13.2	11.0	12.0	14.5	12.4	16.2	14.8	8.1	16.6		
Very Good	33.7	33.9	28.3	35.9	32.7	36.0	32.4	33.9	36.4		
Good	38.1	40.6	45.0	33.6	39.4	34.0	40.8	40.7	30.8	0.257	
Moderate	12.9	12.2	12.8	14.1	13.2	11.5	10.0	14.9	14.6		
Bad	2.1	2.4	2.0	2.0	2.4	2.4	2.0	2.4	1.6		
Private Health insurance											
No	84.3	81.5	86.5	82.4	89.2	86.6	82.8	83.5	82.2	0.184	
Yes	15.7	18.5	13.6	17.6	10.8	13.4	17.2	16.5	17.8		
Subjective risk											
Above average	7.6	11.0	7.6	7.0	6.4	4.7	6.4	9.7	8.3		
Average	45.7	45.3	43.0	42.2	46.6	49.4	43.6	49.2	46.3	0.385	
Below average	43.4	41.3	44.2	46.1	43.4	43.5	46.4	39.5	42.7		
Do not know	3.3	2.4	5.2	4.7	3.6	2.4	3.6	1.6	2.8		
Suffered accident											
No	72.7	68.9	72.1	70.7	76.9	73.5	70.4	72.6	76.3	0.427	
Yes	27.3	31.1	27.9	29.3	23.1	26.5	29.6	27.4	23.7		

Table III. WTP in SE and JE (winsorized[a])

Evaluation Mode			S1	S3	S4	S6
SE		Mean	181.3	199.2	289.1	436.7
		Median	50	50	101	100
		N	254	256	251	251
JE	Group 5	Mean	72.7			
		Median	11			
		N	253			
		SE vs. JE (t-test p-value)	0.000			
		(Mann-Whitney p-value)	0.000			
	Group 6	Mean		117.2	226.6	
		Median		32.5	65	
		N		250	250	
		SE vs. JE (t-test p-value)		0.000	0.0762	
		(Mann-Whitney p-value)		0.003	0.015	
	Group 7	Mean		183.5	419.7	688.4
		Median		60	100	200
		N		248	248	248
		SE vs. JE (t-test p-value)		0.525	0.017	0.003
		(Mann-Whitney p-value)		0.383	0.209	0.000
	Group 8	Mean	84.1			584.4
		Median	30			150
		N	251			253
		SE vs. JE (t-test p-value)	0.000			0.070
		(Mann-Whitney p-value)	0.000			0.005
	Total	Mean	78.4	150.2	322.8	635.9
		Median	20	50	100	160
		N	504	498	498	501
		SE vs. JE (t-test p-value)	0.000	0.012	0.427	0.009
		(Mann-Whitney p-value)	0.000	0.229	0.497	0.000

Sensitivity to scope within EM (p-values)		SE		
		S1	**S3**	**S4**
SE	**S3**	0.495		
	S4	0.000	0.005	
	S6	0.000	0.000	0.0088[b]
		JE		
JE	**S3**	0.000		
	S4	0.000	0.000	
	S6	0.000	0.000	0.000

a. We substitute the value of the 12 highest observations with the value of the 13th highest observation.
b. Not significant at 5% level using Mann-Whitney

Table IV. Differences within JE (t-tests)[a]

		S1		S2		S3		S4		S5		S6		S7	
		G5	G8	G5	G8	G6	G7	G6	G7	G7	G8	G7	G8	G5	G6
S2	G5	0.000	**0.069**												
	G8	0.000	0.000												
S3	G6	0.003	0.024	**0.603**	**0.099**										
	G7	0.000	0.000	0.000	**0.148**										
S4	G6	0.000	0.000	0.000	0.008	0.000	**0.139**								
	G7	0.000	0.000	0.000	0.000	0.000	0.000								
S5	G7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000						
	G8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.650**						
S6	G7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005				
	G8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.042	**0.765**	0.000				
S7	G5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.413**	**0.311**	**0.710**	0.017	**0.206**		
	G6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.009	**0.307**	0.025	**0.762**	**0.469**		
S8	G5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	**0.113**	0.008	**0.713**	**0.191**	0.000	**0.538**
	G6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.031	0.001	**0.350**	0.061	0.003	**0.609**

a. Shadowed cells correspond to within-subjects comparisons.

Table V. Error ($WTP_i > WTP_j$ for $i<j$) and non-discrimination –ND- ($WTP_i = WTP_j$ for $i<j$) (%)

	S1		S2		S3		S4		S5		S7	
	Error	ND	Error	ND	Error	ND	Error	ND	Error	ND	Error	ND
S2	2	48.7										
S4					1.8	45.0						
S5	1.2	28.7	1.6	35.1	1.2	31.0	4.0	50.4				
S6	1.2	28.3	1.2	32.0	1.2	25.8	2.0	43.1	2.8	50.3		
S7	1.2	26.1	0.8	30.5	1.6	28.4	2.0	35.2				
S8	0.8	25.6	0.0	27.4	1.6	25.2	1.6	31.2			2.4	56.6

Appendix. Non Fatal Road Injuries (NFRIs) for valuation

F (S1)	W (S2)
• Does not require hospitalisation, the patient is treated in outpatient settings **After Effects:** • Mild to moderate pain for 1 week • There are difficulties in work and leisure activities that gradually reduce • After 3 or 4 months, full recovery without any sequelae	**In hospital:** • 1 week • Mild pain **After Effects:** • Pain or discomfort for several weeks • There are difficulties in work and leisure activities that gradually reduce • After 3 or 4 months, full recovery without any sequelae
X (S3)	**V (S4)**
In hospital: • 2 weeks • Moderate pain **After Effects:** • Pain gradually reduces • There are difficulties in work and leisure activities that gradually reduce • After 18 months, full recovery without any sequelae	**In hospital:** • 2 weeks • Moderate pain **After Effects:** • moderate to severe pain for 1-4 weeks • Then, the pain gradually fades, but reappears when performing certain activities • There exist permanent restrictions to work and leisure activities
S (S5)	**R (S6)**
In hospital: • 4 weeks • Moderate to severe pain **After Effects:** • moderate to severe pain for 1-4 weeks • Then, the pain gradually fades, but reappears when performing certain activities • There exist permanent restrictions to work and leisure activities	**In hospital:** • More than 4 weeks, possibly several months • Moderate to severe pain **After Effects:** • Lifelong chronic pain • There are major and permanent restrictions to work and leisure activities • Possibly some prominent and permanent scars
N (S7)	**L (S8)**
In hospital: • More than 4 weeks, possibly several months • Inability to use the legs and arms, possibly due to paralysis or amputation **After Effects:** • Confined to a wheelchair for the rest of life • Dependent on others for many physical needs such as dressing and toileting	**In hospital:** • More than 4 weeks, possibly several months • Head injuries that cause permanent brain damage **After Effects:** • Mental and physical abilities greatly reduced for the rest of your life • Dependent on others for many physical needs such as dressing and toileting

Note: S1 was shown as F to the subject, S2 as W, S3 as X, S4 as X, S4 as V, S5 as S, S6 as R, S7 as N, S8 as L.