External validity of WTP estimates: comparing preference and WTP-space model results

Romain Crastes[a *], Olivier Beaumais[b, c], Pierre-Alexandre Mahieu[d], Pablo Martínez-Camblor[e, f], Riccardo Scarpa[g, h]

[a]Agri'terr LECOR, ESITPA

3 rue du Tronquet, 76134 Mont Saint Aignan, France

[b]EconomiX, UMR CNRS 7235, Paris West University Nanterre La Défense

Bâtiment G, 200 Avenue de la République, 92001 Nanterre cedex, France

[c]LISA, UMR CNRS 6240, University of Corsica Pasquale Paoli

Avenue Jean Nicoli, BP 52, 20250 Corte, France

[d]LEMNA, EA-4272, University of Nantes

Chemin de la Censive du Tertre, 44322 Nantes, France

[e]OIB-FICYT, Biosanitary Research Bureau

Calle Matemático Pedrayes, 25, Entresuelo, 33005 Oviedo, Asturias, Spain

[f]Oviedo University

Calle San Francisco, 1, 33003 Oviedo, Asturias, Spain

[g]Gibson Institute, Queens University, Belfast

Mediacal Biology Centre, Lisbum Road, Belfast BT9 7BL, United Kingdom

[h]Department of Economics, University of Waikato,

Private Bag 3105, Hamilton, New Zealand

* Corresponding author at : Unité Agri'terr, ESITPA, 3 rue du Tronquet, 76134 Mont Saint Aignan, France. Tel.: +33 6330 47867 ; fax : +33 2350 52740; email : rcrastes@gmail.com

External validity of WTP estimates: comparing preference and WTP-space model results

Abstract

We introduce a protocol for measuring the external validity of competing Willingness-To-Pay (WTP) distributions derived from Random Parameter models for a given set of Discrete Choice Experiment (DCE) data. This protocol is illustrated by comparing two recent advances in the field of choice modeling: the cost-income ratio in preference space approach and the willingness-to-pay space approach. The protocol is based on a two-round survey. Round one consists in a standard DCE survey at the end of which different competing models are estimated. Round two introduces new respondents from the same survey area. In addition to the DCE survey, these new respondents are asked to repeatedly choose, between a set of values randomly drawn from the competing models previously estimated, the one closest to their true preferences. Respondents finally state the interval that reflects their true preferences. An external validity criterion is then obtained by using a new non-parametric test based on the common area of kernel density estimates. Results indicate that 72 % of the respondents prefer values from the willingness-to-pay space model. Moreover, test results indicate that the willingness-to-pay distribution derived from this model is 1.72 times closer to respondent's true preferences in comparison to the preference space model.

1. Introduction

Recent efforts in choice modeling have focused on the development of econometric tools that accurately represent heterogeneity in taste across choice makers (Train, 2009, Fiebig *et al*., 2010). More precisely, random coefficient models have almost become common place in the field of monetary valuation based on choice experiments. This category of models includes the mixed logit and the generalized multinomial logit, both of which provide ways to derive WTP distributions rather than point estimates. However, implied WTP distributions may greatly vary depending on model specifications. Competing modelling approaches often lead to competing WTP distributions, which may question their validity. This problem is, for example, well-illustrated in the recent debate in the literature on the relative merits of specifications based on utility parameterizations in the WTP space instead of the more conventional specifications in preference space.

Indeed, in their seminal paper on the WTP space approach, Train and Weeks (2005) declared that the WTP distributions they derived from models in preference space translate “untenable” implications because of their “unreasonably” large variance in comparison to the WTP distributions derived from models in WTP space. However, the preference space method was found to fit the data better than the WTP space approach. The authors conclude that alternative distributional specifications are required to either provide more reasonable WTP distributions in preference space or better fit the data in WTP space. Sonnier *et al*. (2007) also found out that the model they estimated in WTP space (referred to in their paper as the “surplus model”) results in more “reasonable” estimates of the distribution of WTP. However, the in-sample fit statistics between the model in preference space and the model in WTP space were found to be “ambiguous”. They conclude that the core of the inferential problems concerning choice experiments revolves around “*whether to implement prior knowledge about WTP […] especially when the data are better fit with arbitrary large WTP values*”. On the contrary to the previous studies, Mabit *et al*. (2006) clearly highlighted the appeal of the WTP space method in comparison of the preference space method. Indeed, the results of their study suggest a better model fit for the data estimated in WTP space as well as more plausible WTP distributions (*i.e*. lower WTP values). A similar result has been obtained by Scarpa *et al*. (2008), who compared the approach in preference space and the approach in WTP space to data on site choice in The Alps. Further merits of the WTP space approach in terms of direct testing of WTP distribution in the estimation stage are discussed in Thiene and Scarpa (2009), while Daly et al. (2012) suggest that WTP space models might be the way forward to overcome the stringent limitations that surround the issue of finite moments of implied WTP distributions from mixed logit models with utility in preference space. Finally, the results obtained by Hole and Kolstad (2012) highlight how WTP estimates might have features with marked differences between model parameterizations.

As shown through this review, the WTP space approach has been argued to produce more reasonable WTP distributions than the preference space approach using the same data. This despite the latter might provide better fit on sample data. The choice of a specification over another is mostly based on expert knowledge. At present the state of practice in choice experiment surveys does not allow researchers to assess to what extent competing WTP distributions *actually* reflect the true preferences of the population. This calls for a rigorous debate on what criterion to use to measure the external validity of model results in terms of WTP distributions. Beyond the specific case of WTP *versus* preference space utility specifications, measuring the external validity of WTP distribution is of interest to the broader stated preference literature. Specifically, whenever different distributional assumptions of attribute coefficients or other modeling issues affect the implied WTP distribution estimates. Revealed preference approaches may also be concerned as, for example, WTP space models are also applied used in travel cost applications (Scarpa, Thiene and Train 2008; Thiene and Scarpa 2009).

In this paper, we propose a new survey methodology for measuring the external validity of WTP distributions derived from random coefficient models. We compare results from random utility choice models with utility in preference- and in WTP-space using data from a CE on the management of erosive runoff-events in a severely flood prone watershed of France. Out protocol consists in conducting surveys in two rounds, with different respondents for each round. The first round consists in a classic CE survey. The second round survey contains an additional set of questions based on WTP distributions obtained using the first round data. The WTP distributions obtained from both rounds are finally compared using a K-Sample test. This is based on the common area of kernel density estimators which, to our knowledge, has never been applied in the field of choice modeling and monetary valuation. The test results give a measure of the external validity of the WTP distributions derived from competing utility specifications. The results from the specification that scores best are suggested to be used for policy making. Although in this paper we do apply our protocol to the comparison of WTP distributions stemming from preference and WTP-space models, the scope of our protocol is much wider and could be applied to measure the external validity of a wide range of model results for a given dataset.

The remainder of this paper is organized as follows. Section 2 describes the specifications in preference and WTP-space. Section 3 presents the two-round survey methodology. Section 4 presents the data and the results. Section 5 discusses the possibility to implement the methodology we suggest in future researches and concludes.

2. Specification

2.1. The preference space approach

The preference space approach refers to the standard model parameterization in DCE. Following the notation proposed by Train and Weeks (2005), the utility a decision-maker *n* derives from choosing the alternative *j* from the choice-set *t* is a function of the cost *p* and a set of non-cost attributes *x:*

$$U_{njt} = \alpha_n p_{njt} + \beta_n{}' x_{njt} + \varepsilon_{njt} \tag{1}$$

α and *β* are randomly distributed and vary over decision makers following a given distribution in order to represent that different people have different tastes, cognitive abilities, etc., and *ε* is a random term which is distributed extreme value with a variance that can vary over decision-makers equal to $\mu_n^2(\pi^2/6)$ with μ_n as the scale parameter for the nth decision-maker. The error term can have the same variance for all decision-makers without affecting the behavior by dividing (1) by the scale parameter:

$$U_{njt} = -\lambda_n p_{njt} + c'_n x_{njt} + \gamma_{njt} \tag{2}$$

$\lambda_n$corresponds to α_n/μ_n and c_n to β_n/μ_n. γ is here IID extreme value with a constant variance equal to $(\pi^2/6)$. Equation (2) corresponds to the model in preference space.

The WTP for a given attribute is obtained through the ratio $w_n = c_n/\lambda_n$. As w_n may also be specified as random according to adequate distributions, the WTP distribution for a given attribute greatly depends on the distribution used for the cost coefficient λ. Holding the cost coefficient fixed may be considered as behaviorally implausible as it implies that there is no heterogeneity in cost sensitivity. On the other hand, common distributions such as the normal, truncated normal, uniform and triangular distributions may prevent the underlying WTP distribution to have finite moments (Daly *et al*. 2012). As a result, the negative of the cost coefficient is often specified to be log-normally distributed as it constrains it to be negative and may allow to obtain distributions of WTP with finite moments under the conditions discussed in Daly *et al*. (2012). However, such specification causes expected WTP values to 'explode', *i.e.* to reach extremely high values because the distribution of λ_{n} makes this coefficient likely to be very close to zero, and hence WTP very large. This problem may be circumvented by using a cost-income ratio variable.

2.2. The cost-income ratio approach in preference space

Indeed, in a recent paper, Giergiczny *et al*., (2012) proposed to introduce a cost-income ratio, an interaction variable constructed by dividing the cost variable by respondent's income in order to prevent WTP values from 'exploding' when the negative of the cost coefficient is log-normally distributed. Following this specification, the WTP for a given attribute is obtained through the ratio $w_n = c_n/(\lambda_n + \lambda_{cost-income}/income_n)$. Because λ_n is expected to be negative by construction, the interacting variable has to be negative as well in order to move the denominator of w_n away from zero. By construction, the cost-income ratio ensures this property because, for normal goods, economic theory assumes that respondents with higher income have higher WTP. As a result, λ_n is moved away from zero for every respondent, although this effect is lower for respondents with higher income. This approach has not been applied elsewhere to our knowledge. A growing alternative to the preference space approach is the WTP space approach, which consists in specifying WTP distributions prior to the model estimation stage rather than deriving them from utility coefficient distributions.

2.3. The Willingness To Pay space approach

The WTP space approach, suggested by Train and Weeks (2005), is based on the idea that convenient distributions for utility coefficients do not imply convenient distributions for WTP and the opposite holds as well. Equation (2) is reformulated so that WTP distributions are directly specified and not derived from utility coefficient distributions:

$$U_{njt} = -\lambda_n p_{njt} + (\lambda_n w_n) x_{njt} + \varepsilon_{njt} \quad (3)$$

where $w_n = c_n/\lambda_n$,

It is worth noting that (2) and (3) are behaviorally equivalent. More precisely, any distribution of λ_n and c_n in equation (2) involves a distribution of λ_n and w_n in (4) and the opposite is also true (Scarpa *et al.*, 2008). The coefficients estimated in WTP space can be estimated using Bayesian techniques or through maximum simulated likelihood and can be interpreted directly as marginal WTP estimates (Train and Weeks, 2005). On the same data, this specification has been shown to produce lower estimates of mean WTP in comparison to those produced using the conventional preference space specifications, which questions the validity of the WTP distributions derived from each of these competing approaches in comparison with respondents' preferences. As a result, we introduce in the next section a new survey protocol to evaluate the external validity of competing WTP distributions. We then provide an empirical application of the proposed protocol.

3. Methods

3.1. Survey protocol

Our protocol relies on a two-round survey design. In this section, we explicit step by step the survey procedure for each round.

3.1.1. Round one

The first round simply consists in gathering enough data for model estimations using a classic choice experiment survey design. The usual recommendations on sample size and fit with census data apply. The questionnaire used during round one is referred to as the base questionnaire in the remainder of the paper. The data from round one are then used for estimating *m* competing models. Indeed, higher values for *m* may result in a higher cognitive burden for the respondents interviewed during phase two (although it also depends on the objective of the researcher). Future applications of the two-round survey design for comparing model results may help defining a more adequate rule for the value of *m*.

3.1.2. Round two

Round two consists in an updated questionnaire with new respondents interviewed from the same population as round one and should be organized as follows:

(i) At first, the second round consists in the administration of the base questionnaire used during round one. A new choice task is then introduced at the end of the base questionnaire. This additional choice task requires the use of a computer operated by the interviewer to be performed. At first, respondents must be asked to indicate, among all the alternatives they chose during the base questionnaire, the one they prefer the most[1]. This information as well as information on respondents' choices and income are then entered in a very simple computer program coded in Visual Basic using Microsoft Excel. The main interface of the computer program is described by Figure 1.

[Figure 1 about here]

(ii) Secondly, the interviewer must use the computer program to draw several positive WTP values for the respondents' favourite alternative based on the random parameter estimates from the *m* models estimated using data from round 1. More precisely, for each model, *k* WTP values are drawn. *k* may vary depending on *m* and with consideration to the cognitive burden faced by the respondents as they will be asked to complete several choice tasks using these values. For example, in the case where $m = 2$, an appropriate value may be $k = 10$. Figure 2 illustrates the WTP calculation interface for two models named 1 and 2.

[Figure 2 about here]

[1] Respondents who select the *status quo* as their favorite alternative cannot be considered as their WTP cannot be computed using standard techniques. Future applications may circumvent this limit.

(iii) The respondent must then choose, among all the WTP values drawn during step (ii), the ones that correspond the most to their true preferences. In this example, we consider the case where k=10 and m=2 with two models m1 and m2 (see Figure 2 for an illustration). At first, the interviewer must report one WTP value drawn from m1 and one WTP value drawn from m2 to the respondents, which corresponds to the first line of the computer interface reported in Figure 2. The respondents must choose which one of the two values is the closest to their preferences. This procedure is then repeated k-1 times with each of the remaining WTP values. Step (iii) is skipped when $m = 1$ for obvious reasons.

(iv) Finally, respondents have to state, among the k WTP values they chose, the one that corresponds to the minimum they are sure to pay and the one that corresponds to the maximum amount they are sure to refuse to pay. The distribution of the WTP values entering the interval selected by the respondents during step (iv) reflects the distribution of the true underlying WTP of the surveyed population.

It is worth noting at this point that the choice task performed during step (iv) echoes the choice task performed in a contingent valuation survey based on the payment ladder format. The value for k should hence be chosen with respect of the literature on payment ladder. It should not be too low to truly represent the range where respondents' preferences may be located but not exceed the number of cells commonly used for designing a payment ladder, which is around 20 (see for example Rowe *et al*., 1996). For example, if $m = 2$ (two models) but $k = 50$ (fifty choice tasks), the respondents may not be able to carry the steps (iii) and (iv) because of the cognitive burden. However, if both k and m equal 2, the resulting sample of WTP values falling in the interval selected by the respondents may not truly represent the range within which respondents' preferences are located.

3.2. External validity measure

3.2.1. Preliminary results

Descriptive statistics on the sample of WTP values chosen by the respondents during step (iii) and the sample of WTP values comprised in the Interval selected by the respondents during step (iv) provide preliminary results. The data obtained from step (iii) may indicate whether respondents majorly chose values drawn from m1 or from m2. Moreover, data from step (iv) provide information on whether the WTP values derived from m1 and m2 majorly correspond to the respondents' true underlying WTP distribution or not. Graphical representations of competing WTP distributions may also provide a clearer insight on respondents' choices. However, these indicators are not sufficient to achieve a precise assessment of the validity of the WTP distributions derived from each of the models considered. Indeed, some models may have a higher chance to provide extreme WTP values, which results in fewer values selected during step (iii) for these models. However, the shape of the WTP distribution derived from these models may still be closer to the shape of the distribution of the WTP values comprised in the interval selected by the respondents during step (iv). in comparison to other models. As a result, we propose for each model to measure the distance between the WTP distribution predicted by the model and the distribution of the WTP values comprised in the interval selected by the respondents during step (iv).

The distance between two distributions may be measured using non-parametric tests. Common tests include the Kolmogorov-Smirnov, Crámer-von Mises and Anderson-Darling tests (Martínez-Camblor *et al*., (2008)). Each of these tests provides a measure of the distance between the two distributions tested (which are, in the same order, the D-stat, the w2-stat and the W2-stat). However, the interpretation of these measures is not sufficiently straightforward to reveal how the competing WTP distributions rank in terms of respective closeness to the distribution of the WTP values comprised in the interval selected by the respondents during step (iv). Moreover, these tests have been shown to have a lower performance in comparison to a new k-sample test based on the common area of kernel density estimator as introduced below (Martínez-Camblor *et al.*, (2008)). This test has never been used in the field of choice modelling to our knowledge and we propose to use it here.

3.2.2. Common area of kernel density estimators test

The common area of kernel density estimators test allows to measure the distance between k-random variables with densities $f_{1,}...,f_{k}$. It consists in measuring each distance among the $\hat{f}_{n1},\ldots,\hat{f}_{nk}$ kernel estimators considered, which defines a test statistic for the null hypothesis $H_0 : f_1 = ... = f_k$ (respectively $H_1 : f_1 \neq ... \neq f_k$). The distance used is the area under the kernel density estimators, which is common to all of them. This area is designed as the common area (*AC*) criterion, introduced and studied by Martínez-Camblor *et al.*, (2008). It is defined by Equation (4):

$$AC = \int \min\left\{ f_1(t),..., f_k(t) \right\}\ dt \qquad (4)$$

[Figure 3 about here]

Figure 3 is a simple illustration of the common area criterion. The figure describes the kernel density estimates of three simulated distributions *f*, *g* and *h*. The grey area is common to all of these three densities and simply corresponds to the *AC* criterion. The *AC* criterion ranges between 0 (absolute discordance) and 1 (absolute concordance).

The direct *AC* estimator is the results of replacing the usually unknown density functions by appropriate estimators, namely kernel density estimators. Its behaviour has been analysed with several simulation studies whose results are available in Martínez-Camblor *et al.*, (2008) and Martínez-Camblor and De Uña-Álvarez (2009). The bootstrap method is usually used in order to approximate *p*-values. The amount of smoothing used for the computation of the kernel density estimators influences the power of the test. The results from the studies previously mentioned suggest that the *AC* test perform better than previous *k*-sample tests, such as Kolmogorov-Smirnov, Crámer-von Mises and Anderson-Darling tests, whenever the smoothing degree is optimally chosen (see Table 2 and Table 3 from Martínez-Camblor *et al.*, (2008) for a deeper insight). The *AC* criterion test has been shown to be useful for studying whether the involved densities present differences in shape or in spread, especially for homogeneous sample sizes. In the context of the two-round survey protocol, the *AC* criterion test is used to measure the distance between the WTP distribution derived from each of the competing models and the distribution of the WTP values entering the intervals selected by the respondents during step (iv), in which case k = 2. Martínez-Camblor *et al.*, (2008) demonstrate that in the two samples case, the *AC* criterion corresponds to:

$$AC = 1 - \frac{1}{2}\int \left| f_1(t) - f_2(t) \right| \, dt \tag{5}$$

This test allows researchers to obtain a measure of the common area, in percentage, between the WTP distribution derived from each of the competing models and the distribution of the true underlying WTP of the surveyed population. Such measure provides a direct comparison between competing models thereby facilitating the identification of the most appropriate model. Moreover, the interpretation of this measure is very straightforward in comparison to the tests previously mentioned. These features make the *AC* criterion well suited for measuring the external validity of competing WTP distributions. The model which provided the WTP distribution for which the *AC* criterion is the highest should be used for policy making.

In the next section, we illustrate the two-round survey protocol for measuring the external validity of WTP distributions with an empirical example. We compare results from a utility specification in preference space and one in WTP space. We use data from a DCE on the management of erosive runoff-events in a severely flood prone watershed of France.

4. Survey

4.1. Round 1

The first round of the DCE on erosive runoff events mitigation measures took place in fall 2011 in the Vallée du Commerce (Upper-Normandy, France) which is a watershed severely exposed to erosive runoff events (floods, mudslides, landslides). The DCE survey is fully described in Crastes *et al.*

(2013). Respondents were asked to state their preferences about a program for preventing and reducing erosive runoff events. Three non-monetary attributes were considered. Each attribute corresponds to a set of broadly described measures aimed at preventing and reducing erosive runoff events. Each attribute could take two levels, « *yes, the measures are included in the program* » or *« no, the measures are not included in the program* ». The first management policy attribute, referred to as *agriculture*, consisted in implementing responsible water management measures in farming production. The second management policy attribute, *infrastructure*, consisted in implementing protective infrastructures, which comprises hydraulic works (absorbing parking, permeable roads) and heavy structures (dams and dikes). Finally, the third management policy attribute, *communication*, corresponded to a set of measures aimed at increasing the general public awareness about erosive runoff events and the measures individuals may take by themselves in order to mitigate these risks. Excluding the status quo, the cost attribute was the only monetary attribute and could take three levels: €12.50, €25 and €37.50. The values taken by the cost attributes have been decided together with the district authorities using the results from public surveys[2] as well as a pre-test survey. More precisely, the cost levels were designed as follows. Firstly, the lower level (€ 12,50) corresponds to the minimum payment under which the policies valued could not be implemented according the watershed district authorities. This level has hence been chosen in order to propose realistic choices[3]. The two other cost levels were chosen to be equidistant in space with the minimum price level to maintain orthogonality in the design. The three cost levels correspond to three options, "low", "medium" and "high" for the increase of the local tax set to be the payment vehicle, which reflects the actual and usual choice context from the watershed district authorities point-of-view. Pre-tests did not report any specific problem regarding such design.

[2] The river basin district authorities of the Vallée du Commerce regularly consults the local population on their concern regarding the watershed management. Such consultations are named public survey and are similar to what is identified as focus groups in the field of Discrete Choice Experiments (DCE). A series of consultations has been carried out while the survey.

[3] The final sample shows that less than 5% of the respondents declared to be in favor of the program but considered the minimum price level as too high. A similar proportion has been found during pre-tests, which indicates that the minimum range for the price attribute has been adequately selected.

The DCE has been designed as follow: 24 alternatives were generated following an orthogonal optimal in the difference factorial design (Street and Burgess, 2007). Beyond the *status quo*, each respondent had to face a second alternative specifically chosen to be the opposite of the first alternative. More precisely, in alternative A, each attribute level was at the other value to alternative B. The cost level was also set to be different. The 24 choice sets have been divided between 4 versions of the questionnaires composed of 6 choice sets each. These four questionnaire versions have been distributed in a balanced way across the respondents. 341 respondents provided complete answers during round one. Respondents were chosen following quota sampling on age, gender and profession in order to ensure that the survey sample fits with census data. Information on income and localization were also gathered. Table 1 provides descriptive statistics.

[Table 1 about here]

Two random coefficient models were estimated using the data gathered during the first round of the experiment. The first model is specified in Preference Space (PS) following the approach proposed by Giergiczny *et al.* (2012) and is referred to as the PS model while the second model is specified in Willingness to pay Space (WS) and is referred to as the WS model. The coefficients for the non-monetary attributes *agriculture*, *infrastructure* and *communication* were set as random and assumed to be normally distributed for both models while the cost coefficient was set to be log-normally distributed for the PS model. In addition, a cost-income ratio variable enters the PS model only. Table 2 provides estimation results for both models.

[Table 2 about here]

It is worth noting that the mean estimates of the non-monetary attribute coefficients are all positive and significant at the 1% level for both models. Moreover, the cost coefficient and the cost-income ratio of the PS model are both negative and significant as expected. Each of the random coefficients entering the models has a significant standard deviation. The sign of the ASC changes depending on model specification. As it was reported in previous studies (Train and Weeks, 2005 ; Sonnier *et al.*, 2007), the log-likelihood for the PS model is slightly higher than for the WS model, while the estimated means of the marginal WTP distributions are higher for the PS model. The external validity of these results is measured through the second round of the survey.

4.2. Round 2

The second round took place in fall 2012 sampling from the same population sampled in round one and under similar conditions. Respondents were confronted with an updated version of the questionnaire as previously explained:

(i) At first, respondents answered the same DCE survey than during round 1 and stated their favourite alternative.

(ii) Secondly, interviewers used the computer program previously introduced to draw 10 positive WTP values for the respondents' favourite alternative based on the unconditional parameter estimates of the PS model and 10 WTP values for this alternative based on the parameters estimates of the WS model, so $m = 2$ and $k = 10$. Unconditional parameter estimates have been chosen rather than posterior parameter estimates conditional on the respondent's choices because the existing literature on preference space *versus* WTP space focuses on the comparison of unconditional parameter estimates. As a result, we use the two-round survey protocol to bring additional knowledge on a problem which has been extensively treated in the literature.

(iii) Thirdly, respondents faced one WTP value derived from the PS model and one WTP value derived from the WS model, chose which one of the two was the closest to their true preferences and repeated this operation for a total of 10 times.

(iv) Finally, respondents stated, among the 10 WTP values they chose, the one that corresponds to the minimum they are sure to pay and the one that corresponds to the maximum amount they are sure to refuse to pay.

In total 102 respondents stratified according to census data on age, sex and income were surveyed during round two. Table 3 provides descriptive statistics for the second round of the survey sample and Table 4 describes respondents' favourite alternative choices.

[Table 3 about here]

[Table 4 about here]

At first, we compare the distribution of the WTP values drawn from the PS model and the distribution of the WTP values drawn from the WS model during step (ii) using kernel density estimates. The PS model produced a higher mean WTP and its distribution exhibits a longer tail in comparison to the WS model as shown by Figure 4 and Table 5:

[Figure 4 about here]

[Table 5 about here]

Table 6 provides descriptive statistics on the WTP values drawn from the PS and the WS models during step two as well as on the WTP values chosen by the respondents during step (iii) and the WTP values entering the intervals selected by the respondents during step (iv).

[Table 6 about here]

The main result from Table 6 is that about 71% of the values chosen during step (iii) were derived from the WS model (718 choices made) while there are only 298 values (29 %) derived from the PS model. However, only 347 values out of the 1015 WTP values selected enter the interval selected by respondents during step (iv) (34.18%). Out of these 347 selected values, 245 (71.64%) were drawn from the WS model. We then investigate, for each pair of WTP values presented to the respondents, the distribution of the difference between the WTP value derived from the PS model and the WTP value derived from the WS model. The aim is to identify whether the values proposed to the respondents mainly involved extreme or rather obvious choices (for example € 10 *versus* € 1000), which would result in a heavily skewed distribution, or rather balanced choices. Results are reported in Figure 5.

[Figure 5 about here]

Figure 5 shows that the distribution of the difference between presented values shows high densities around zero and a long tail on the right side. This result indicates that most of the values that have been presented to the respondents did not imply rather obvious choices but also that the values presented drawn from the PS model were generally higher to the values drawn from the WS model, which is consistent with Figure 4.

The external validity of the WTP distributions derived from the two competing choice models estimated in the first round of data collection is finally measured using the *AC* criterion test. Figure 6 and 5 report the kernel density estimates of the sample of WTP values drawn from the PS model and the WS model during step (ii) in comparison to the sample of WTP values entering the interval selected by the respondents during step (iv). The *AC* criterion test is used to assess and compare the common area between these distributions. Table 7 reports the results of the test ran over these kernel densities.

[Figure 6 about here]

[Figure 7 about here]

[Table 7 about here]

It is shown by Table 7 that about 38% of the area under the kernel density estimator of the the sample of WTP values entering the interval selected by the respondents during step (iv) is common with the kernel density estimator of the sample of WTP values drawn from the PS model while it is about 67 % for the kernel density estimator of the sample of WTP values drawn from the WS model. Test results give a precise measure of the validity of the WTP values provided by the PS model in comparison to the WS model. According to the *AC* criterion, the WTP distribution drawn from the WS model is about 1.72 (0.631/0.369) times closer to the true underlying WTP distribution of the surveyed population in comparison to the PS model. As previously stated, the model that scores the highest *AC* criterion should be used for policy making. The WTP distributions derived from the WS model have higher external validity and should hence be used for policy making rather than the WTP distribution derived from the PS model, which has comparatively lower validity. It is worth noting that these results may be case-specific and that future applications of the two-round survey design may find different results with this method.

5. Discussion and conclusion

The use of random parameter models is now common place in the field of choice modelling. The wide range of random parameter specification available often leads to competing models that provide substantively different estimates of WTP distributions. This is challenging for the researcher because of the absence of a clear, external measure of the validity of competing WTP distributions. The main contribution of this paper is to propose a new survey protocol for measuring and comparing the external validity of the WTP estimates from different random parameter specifications from a given set of choice data.

The protocol consists in dividing DCE surveys in two rounds. The first survey round is dedicated to data collection as in any other DCE survey. The data collected in this round are used to estimate a set of models. The second survey round is dedicated to collect data that are used to measure the external validity of the WTP distributions derived from the models previously estimated on the data collected in the first survey round. In the second survey round respondents are asked to choose, among a set of WTP values randomly drawn from the distributions of WTPs implied by the round 1 models, the WTP values for their favourite alternative that are closest to their preferences. This operation may be repeated several times with different random values. Finally, respondents have to state, among all the value chosen, the upper and lower bounds that best restrict the WTP interval corresponding to their true preferences for the given alternative. The external validity of competing WTP distributions and hence models is finally measured by using a new k-sample test based on the common area of kernel density estimators. This test provides a measure comprise between 0 and 1of the common area between two or more kernel density estimates, named the *AC* criterion. It is used to compare the WTP distributions derived from each competing model to the distribution of the interval corresponding to respondents' true preferences. The *AC* criterion is used as an external validity measure of a given WTP distribution. It is suggested that the model that obtains the highest *AC* criterion should be used for policy making. In this paper, the two-round survey protocol has been illustrated by comparing the preference space approach with a cost-income ratio introduced by Giergiczny *et al*. (2012) to the utility specification in WTP space approach (Train and Weeks, 2005) using data from a DCE on the mitigation of erosive runoff events in France. The *AC* criterion is equal to 0.67 for the model specified in WTP space while it is 0.38 for the model specified in preference space. The model specified in WTP space is hence suggested to have higher external validity and to be better suited for policy making.

Although we took the specific example of comparing the willingness to pay space approach with the cost-income ratio approach, the methodology we suggest can be extended to every situation where it is necessary to measure the validity of estimated WTP distributions. Moreover, the generalization of this kind of survey procedure would greatly improve benefit transfer and meta-analysis applications as it would allow to select and treat primary studies depending on the external validity of their WTP estimates. Researchers could set up *AC* thresholds below which model results may not be considered as externally valid. Finally, as a caveat we note that choice experiments data are expensive and time consuming to obtain. So, the methodology we propose may be criticised because it requires to set up a second survey round. However, further work on WTP estimates comparison and on external validation of model specifications may consider embedding the approach we suggest within a multi-stage sequential learning approach. More precisely, pre-tests could be extended in order to gather enough observations to estimate reliable models that would be used at the survey stage in order to provide measures of the external validity of WTP distributions, perhaps even by adjusting iteratively the experimental designs of subsequent stages adaptively, following the principles of Bayesian adaptive designs (Scarpa *et al*., 2007; Vermeulen *et al*., 2011) for externally valid specifications.

References

Crastes, R., Beaumais, O., Arkoun, O., Laroutis, D., Mahieu, P.A., Rulleau, B., Hassani-Taibi, S., Barbu, V.S., Gaillard, D., 2014. Erosive runoff events in the European Union: using discrete choice experiment to assess the benefits of integrated management policies when preferences are heterogeneous. Ecological Economics 102: 105-112.

Daly, A., Hess, S., Kenneth, T., 2012. Assuring finite moments for willingness to pay in random coefficient models. Transportation 39(1):19-31.

Fiebig, D.G., Keane, M.P., Louviere, J., Wasi, N., 2010. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. Marketing Science 29(3): 393-421.

Giergiczny, M., Valasiuk, S., Czajkowski, M., De Salvo, M., Signorello, G., 2012. Including cost income ratio into utility function as a way of dealing with ‘exploding’ implicit prices in mixed logit models.Journal of Forest Economics.

Hole, A.R. and Kolstad, J., 2012. Mixed logit estimation of willingness to pay distributions: a comparison of models in preference and WTP space using data from a health-related choice experiment. Empirical Economics 42(2): 445-469.

Mabit, S.L., Caussade, S., Hess, S., 2006. Representation of taste heterogeneity in willingness to pay indicators using parameterization in willingness to pay space. In: Proceedings ETC 2006.

Martínez-Camblor, P., De Uña-Álvarez, J., Corral, N., 2008. K-Sample test based on the common area of kernel density estimators. Journal of Statistical Planning and Inference 138(12): 4006-4020.

Martínez-Camblor, P., De Uña-Álvarez, J., 2009. Non-parametric k-sample tests: Density functions vs distribution functions. Computational Statistics and Data Analysis 53(9): 3344-3357.

Rowe, R. D., Schulze, W. D., Breffle, W. S., 1996. A test for payment card biases. Journal of Environmental Economics and Management 31: 178-185.

Scarpa, R., Thiene, M., Train K., 2008. Utility in willingness to pay space: a tool to address confounding random scale effects in destination choice to the Alps. American Journal of Agricultural Economics 90(4): 994-1010.

Scarpa, R., Campbell, D. Hutchinson, W. G. 2007. Benefit estimates for landscape improvements: sequential Bayesian design and respondents' rationality in a choice experiment study. Land Economics. (November) 83(4):617-634.

Sonnier, G., Ainslie, A., Otter, T., 2007. Heterogeneity distributions of willingness-to-pay in choice models. Quantitative Marketing Economics 5(3): 313-331.

Street, D. J., Burgess, L., 2007. The construction of optimal stated choice experiments: theory and methods. Hoboken, Wiley.

Thiene, M., Scarpa, R., 2009. Deriving and testing efficient estimates of WTP distributions in destination choice models. Environmental and Resource Economics, 44:379–395

Train, K. E., Weeks, M., 2005. Discrete choice models in preference space and willingness-to-pay space. In: Scarpa R, Alberini A (eds) Application of simulation methods in environmental and resource economics. Springer, Dordrecht : 1-16.

Train, K., 2009. Discrete choice models with simulations. Cambridge University Press, New York.

Vermeulen, B., Goos, P., Scarpa, R. and Vandebroek, M. L. 2011. *Bayesian Conjoint Choice Designs to Measure the WTP*. Environmental and Resource Economics, 48:129-149

Table 1

Descriptive statistics - first round.

Variable	Description	Mean	Std. Dev.	Min	Max
Age	Age in years	48.241	16.918	20	92
Female	= 1 if female, 0 else	0.545	0.497	0	1
Income	Income, in thousands of euros per month	2.037	1.08	0.65	3.999

Table 2

Choice modelling results.

Variables	MIXL + cost-inc. Ratio (model PS)			WTP Space model (model WS)		
	Coeff.	Std. Err.	P-value	Coeff.	Std. Err.	P-value
Mean of random parameters						
Cost	-2.624	0.229	0.000			
Agriculture	2.181	0.203	0.000	21.505	2.896	0.000
Infrastructure	2.298	0.186	0.000	23.179	2.491	0.000
Communication	1.373	0.179	0.000	14.126	2.638	0.000
Non-random parameters						
Asc	1.919	0.278	0.000	-5.415	2.374	0.023
cost/income	-0.028	0.014	0.046			
Standard deviations						
Agriculture	1.551	0.301	0.000	33.159	3.261	0.000
Infrastructure	1.231	0.202	0.000	29.672	2.755	0.000
communication	1.26	0.348	0.000	23.231	2.929	0.000
Cost	2.489	0.293	0.000			
log-likelihood			-1118.885			-1295.992

Table 3

Descriptive statistics - second round.

Variable	Description	Mean	Std. Dev.	Min	Max
age	Age in years	49.352	15.004	21	85
female	= 1 if female, 0 else	0.529	0.501	0	1
income	Income, in thousands of euros per month	2.528	1.01	0.65	3.999

Table 4

Break-down of favorite choice.

Choice	Freq.	Percent
Agriculture	6	5.88
agriculture + infrastructure	28	27.45
agriculture + infrastructure + communication	33	32.35
Infrastructure	5	4.9
infrastructure + communication	16	15.69
Communication	2	1.96
communication + agriculture	12	11.76
Total	102	100

Table 5

Model distributions.

Percentiles	1%	5%	10%	25%	50%	75%	90%	95%	99%
PS	0.079	0.652	1.796	8.092	45.977	150.717	319.604	414.859	675.668
WS	1.213	4.389	8.639	19.245	37.85	61.901	90.526	112.187	156.246

Table 6.

Survey results

	Number of values chosen	Mean WTP (in euros per year)	Std. Dev.	min.	max.	Number of values chosen	Mean WTP (in euros per year)	Std. Dev.	min.	max.
WTP values drawn during step (ii)		108.511	150.743	0.01	1132.389		44.831	33.647	0.134	211.281
WTP values chosen during step (iii)	297	24.782	29.389	0.039	207.37	718	36.338	27.523	0.134	200.333
WTP values comprised in the interval selected during step (iv)	97	18.778	12.296	0.098	52.617	245	26.194	12.669	2.409	68.937

Table 7
Common area test between the Distribution of the WTP Values comprised in the interval selected by the respondents during step (iv) (IVD) and the WTP distributions derived from the PS model (PSD) and the WS model (WSD).

Model		*AC*	P-value
PS model	H0: PSD = IVD H1: PSD ≠ IVD	0.369	0.000
WS model	H0: WSD =IVD H1: WSD ≠ IVD	0.631	0.000
Bandwidht values grid = (0.5, 1, 3, 6, 9), α = 0.05			

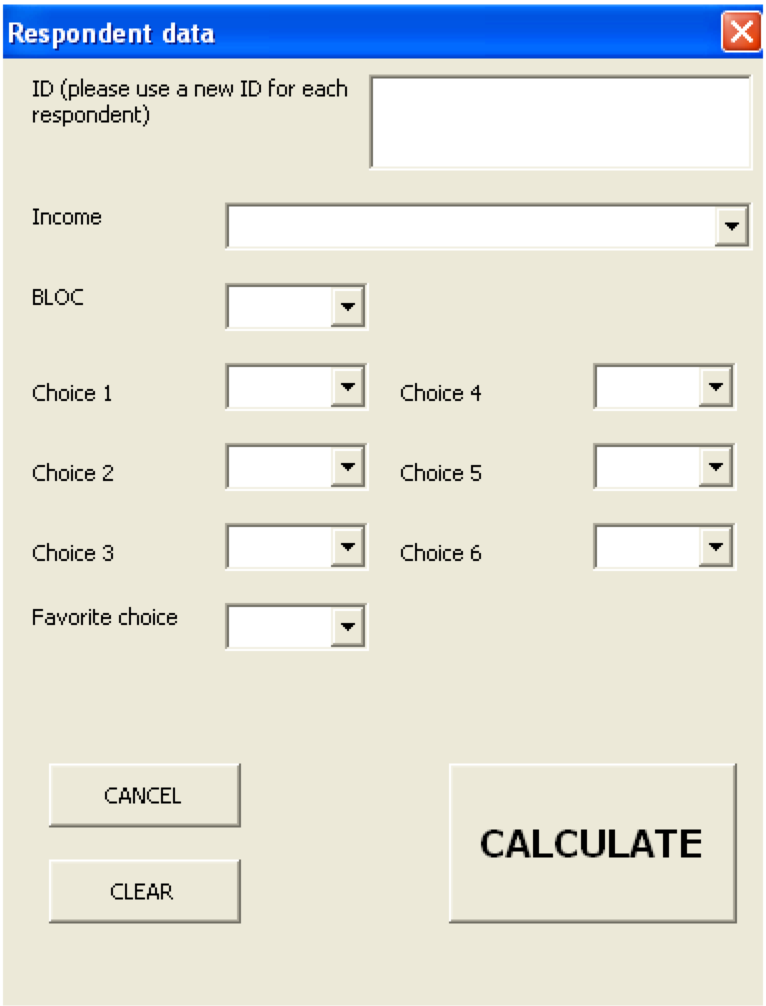
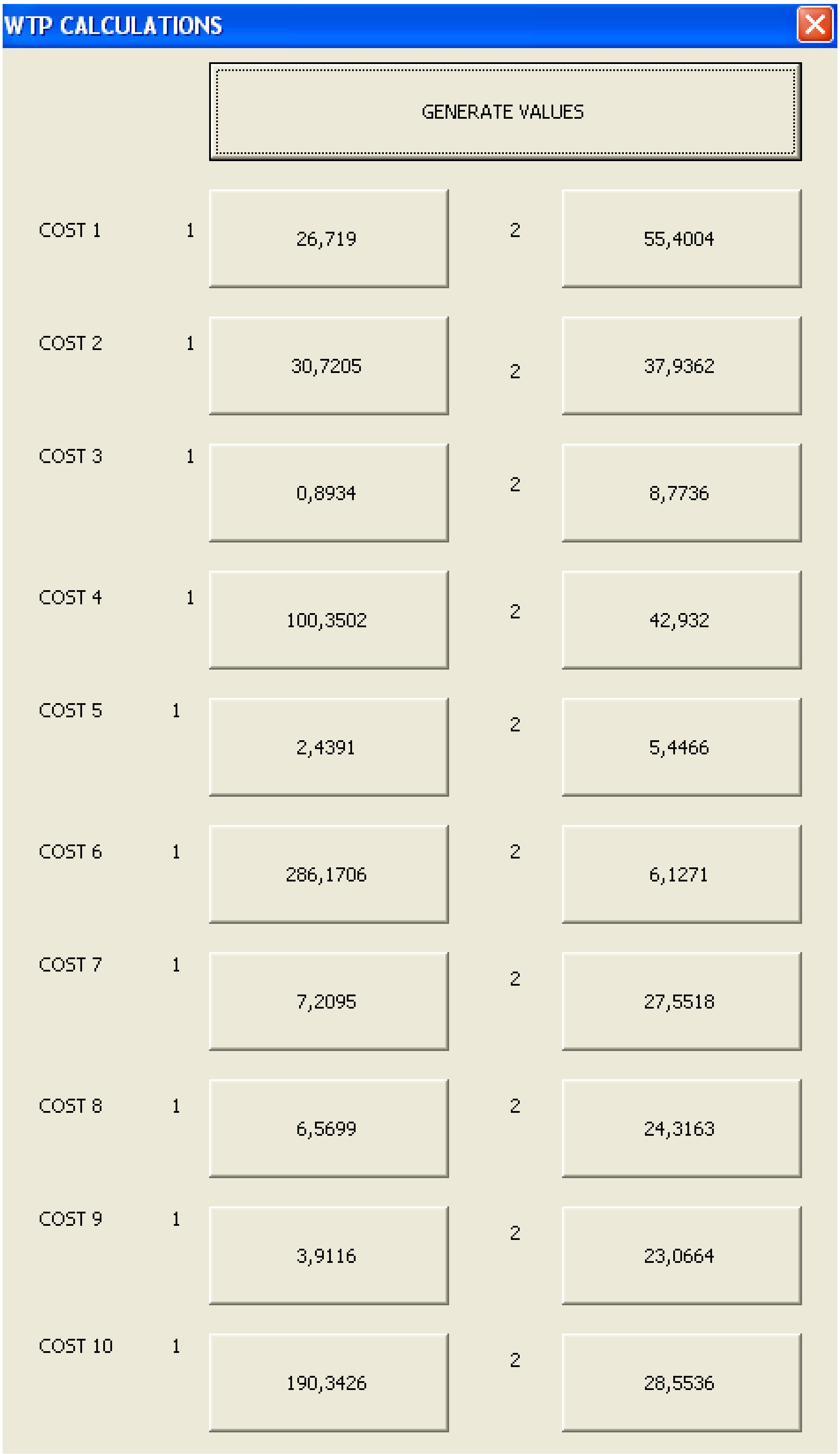


Figure 1. Main interface of the computer program



COST 1	1	26,719	2	55,4004
COST 2	1	30,7205	2	37,9362
COST 3	1	0,8934	2	8,7736
COST 4	1	100,3502	2	42,932
COST 5	1	2,4391	2	5,4466
COST 6	1	286,1706	2	6,1271
COST 7	1	7,2095	2	27,5518
COST 8	1	6,5699	2	24,3163
COST 9	1	3,9116	2	23,0664
COST 10	1	190,3426	2	28,5536

Figure 2. WTP calculation interface of the computer program

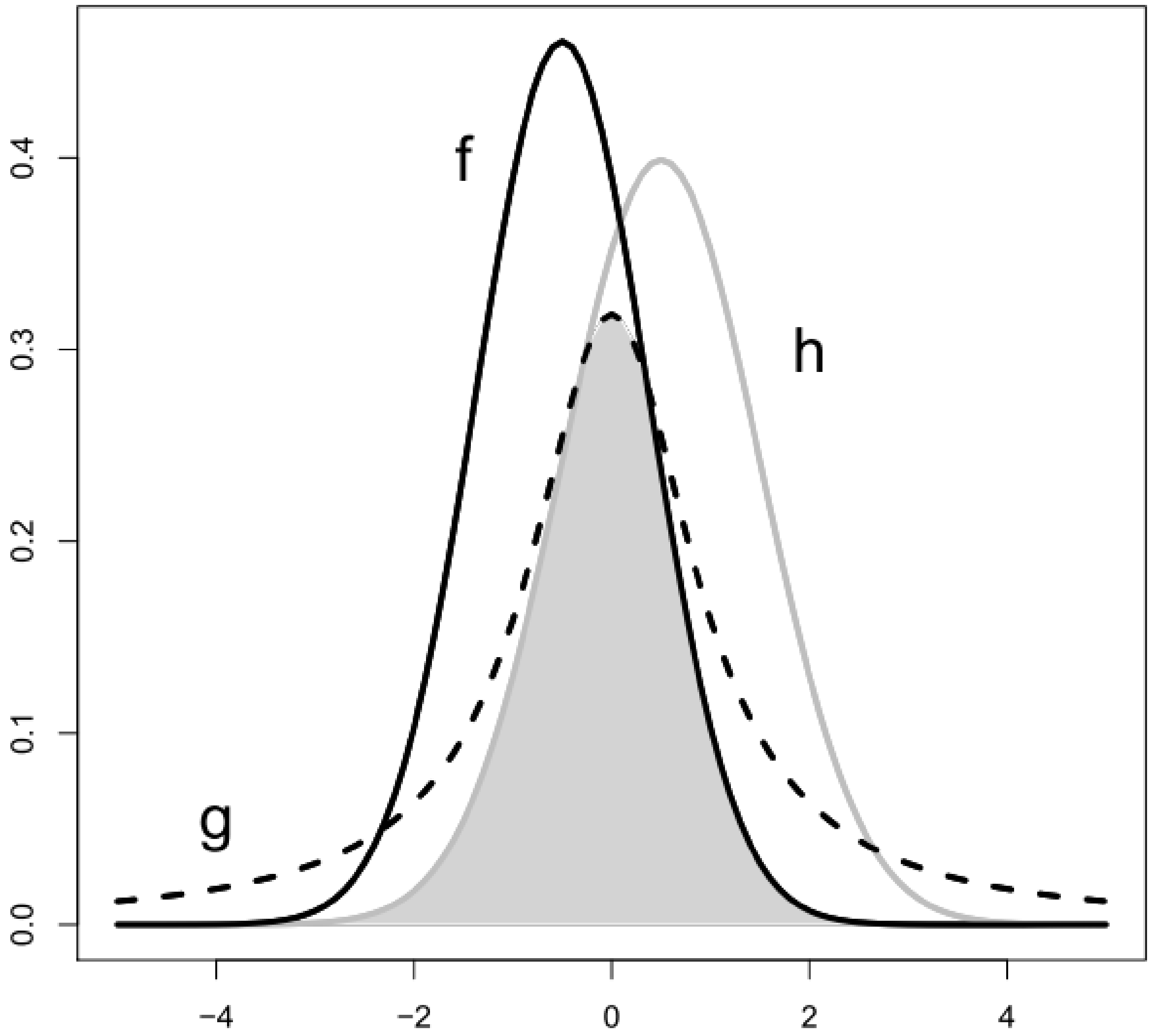


Figure 3. The *AC* criterion test

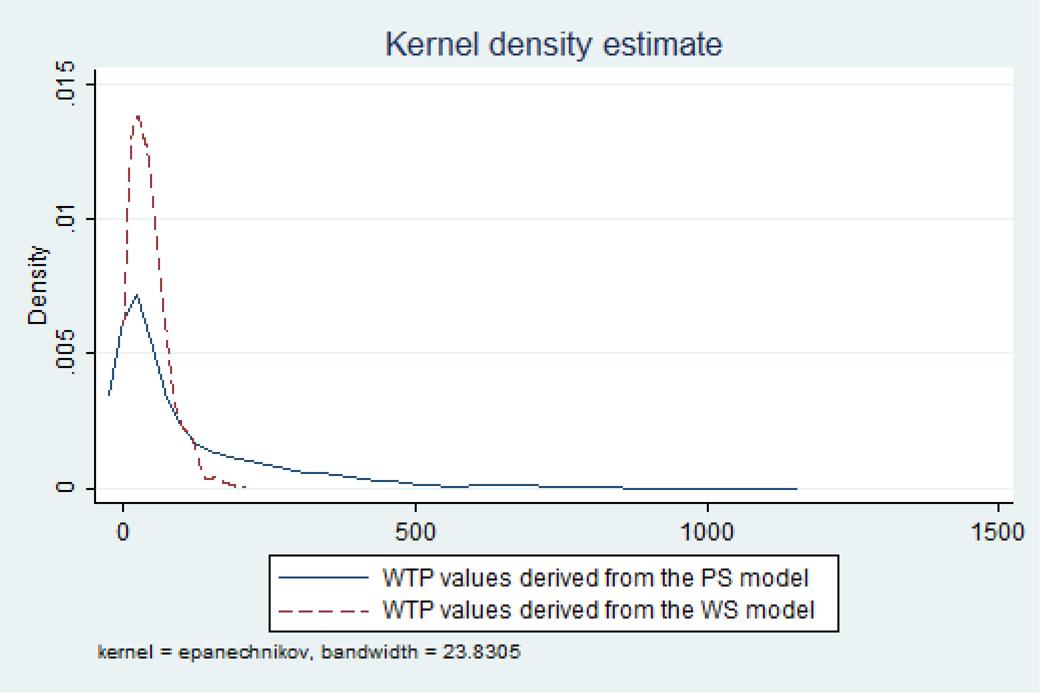


Figure 4. Kernel density estimates of WTP values drawn from the two competing models during step (ii)

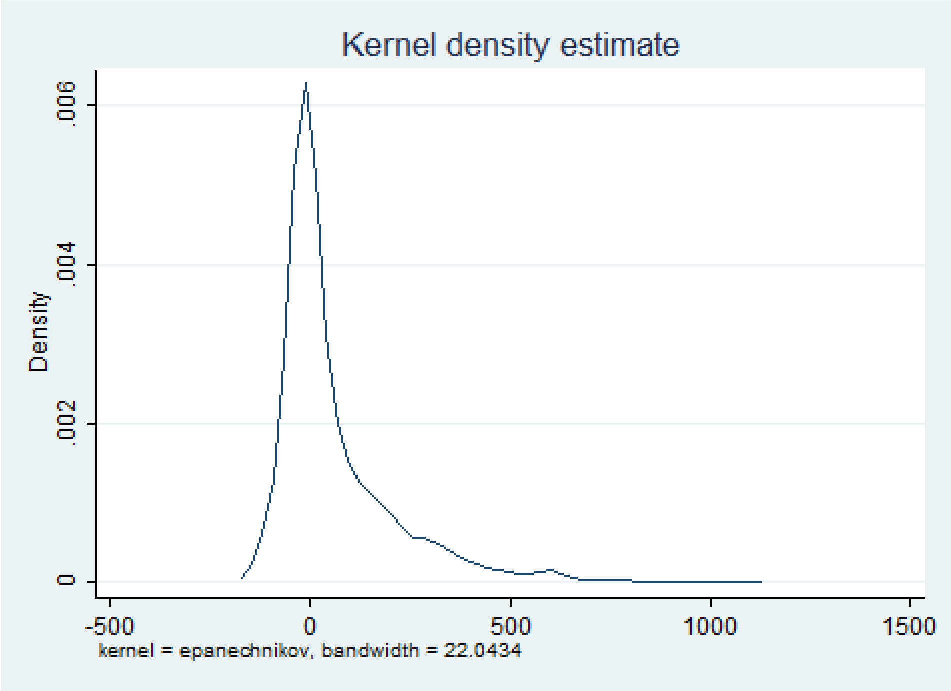


Figure 5. Kernel density estimate of the difference between pairs of WTP values presented to the respondents during step (iii)

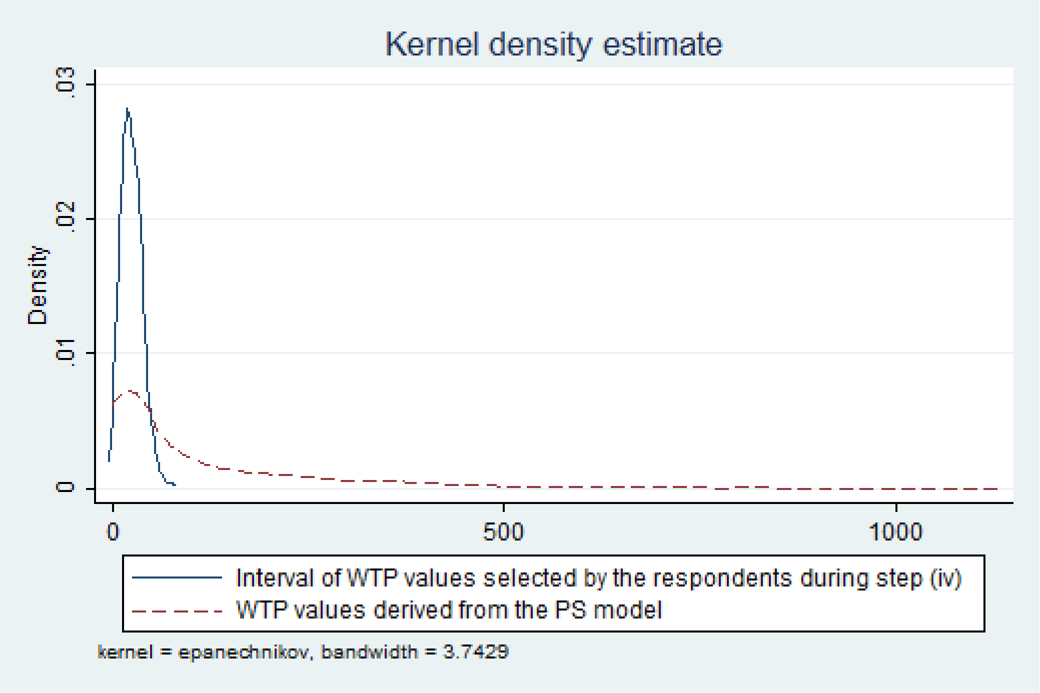


Figure 6. Kernel density estimates of the interval of values selected by the respondents during step (iv) *versus* WTP values derived from the PS model during step (ii)

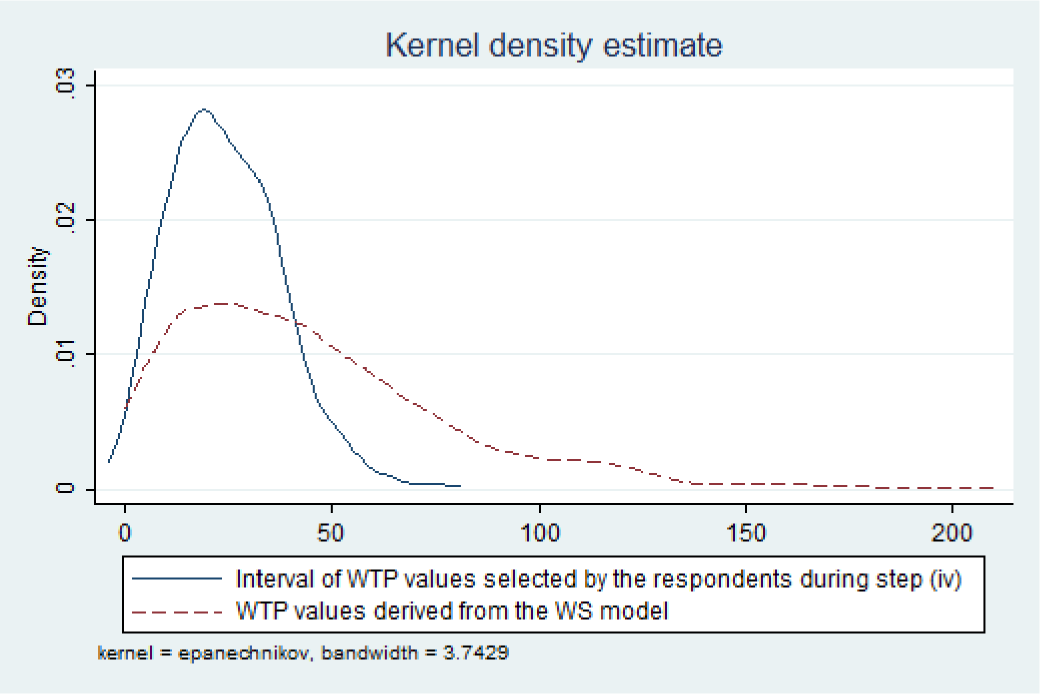


Figure 7. Kernel density estimates of the interval of values selected by the respondents during step (iv) *versus* WTP values derived from the WS model during step (ii)