Do experimental protocols in Conjoint Analysis matter in non Hypothetical settings?

Yangui, A[1*]., Akaichi, F[2]., Costa-Font, M[1]., and Gil, J.M[1].

[1] CREDA-UPC-IRTA "The Center for Agro-food Economy and Development"

[2] Scotland's Rural College (SRUC)

Contributed Paper prepared for presentation at the 88th Annual Conference of the Agricultural Economics Society, AgroParisTech, Paris, France

9 - 11 April 2014

*Corresponding author: ahmed.yangui@upc.edu

CREDA-UPC-IRTA, Parc Mediterrani de la Tecnologia Edifici ESAB, C/Esteve Terrades, 8, 08860 Castelldefels

Abstract

This paper aims at comparing the performance of three conjoint analyses (CA) in terms of estimated partworths, predictive power and estimated WTP: choice experiment (CE); ranking conjoint analysis (RCA) and best-worst scaling (BWS). Comparisons are made in a non-hypothetical setting. For comparison purposes in the last two formats only the information on the most preferred option is considered. The hypothetical CE is used as the benchmark. Olive oil is the food product used in our experiment. Results reveal preferences regularity between samples' responses across the formats implying not statistically differences in the marginal participants' WTP. Moreover, in an incentive compatible context, RCA and BWS compared with CE provide similar results regarding to the in-sample and out-of-sample predictive power and also in terms of decision consistency when just only the first rank data is analyzed.

Keywords: *conjoint analysis, best worst scaling, external validity, experimental economics, hypothetical bias.*

JEL code: C91, D12.

1. Introduction

Since its introduction conjoint analysis (CA) has become one of the most popular marketing research tools (Lusk et al., 2008; Campbell and Lorimer, 2009). In CA's tasks, participants are provided with at least two product profiles and are asked to rate/rank them or select the profile they prefer most. The most widely used CA format to elicit consumers' preferences for market and non-market goods is choice experiment (CE). In CE, respondents are shown a set of combinations of attributes (i.e., profiles) and are asked to indicate which of the combinations or profiles they would purchase. CE gained popularity thanks to its ability to mimic the real market setting where consumers who are faced with competing products purchase the product that fits most their preferences. However, it is informationally inefficient, since it only allows the observation of the most preferred option (Lusk et al., 2008; Louviere et al., 2008; Lanscar et al., 2013). According to Lancsar et al. (2013), there are three ways to gain more insights about individual preferences in CE: 1) increasing the sample size; and/or 2) asking the respondents to evaluate more choice sets; or 3) increasing the number of options per choice set.

In contrast with CE, participants in a ranking conjoint analysis (RCA) are provided with a set of product concepts and they are asked to rank them from the most to the least preferred. The use of RCA as an alternative to CE is becoming popular since it provides information not only about the most preferred product concept but also about consumers' preferences for all the product concepts included in a choice set, which could lead to a more efficient preference estimates (Chang et al., 2009; Louviere et al., 2008; Lusk et al., 2008).

Despite the wide application of the aforementioned CA formats (i.e. CE, and RCA) over the last two decades, few researchers, however, have compared their performance in terms of the estimated marginal partworths, the predictive power of the derived models, and the reliability of the Willingness to Pay (WTP) values deduced from the estimated partworths (Boyle et al., 2001; Holmes and Boyle, 2001; Morrison and Boyle, 2001; Siikamäki et al., 2007). Although, the results of previous studies reveal differences between these formats, all of these studies featured at least one of the shortcomings such as used different experimental design for choice and ranking methods, different number of alternatives, or different sample size. Overcoming some of these shortcomings, for example, Caparros et al., 2008 pointed out that CE and RCA provide similar results, when a similar experimental design is used for both CA formats. Akaichi et al. (2013) confirmed this result for small choice sets (four alternatives); however, they found discrepancies between respondents' preferences in CE and RCA when large choice sets are used.

Recently, Louviere et al. (2004) introduced another CA format named best worst scaling (BWS). The BWS approach consists in asking respondents to firstly choose the best and the worst option, then the second best and the second worst options from the remaining options and so on until a complete preference ordering of all the options is obtained. BWS tasks seem to be easier to handle by respondents due to human skills at identifying extremes (Helson, 1964; Flynn and Marley, 2012). As in the RCA, the additional choice information obtained from BWS has been showed to improve the statistical efficiency of choice models, especially, when it is combined with an appropriate experimental design (Lancsar et al., 2013). Nevertheless, the previous studies, surprisingly, did not assess the comparability of BWS to CE and RCA, although BWS's superiority in terms of realism and ease of its implementation. To fill this gap, this paper stands out by comparing the performance of CE, RCA and BWS in terms of estimated partworths, predictive power and estimated WTP considering, in the estimation of partworths in RCA and BWS when only the information on the most preferred option is considered.

Furthermore, most of the past literature focused on assessing the incentive compatibility of CA and proposing modified CA formats to incentivize subjects to truthfully reveal their preferences (Lusk and Schroeder, 2004; Ding et al., 2005; Ding et al., 2007; Lusk et al., 2008). However, it is worth noting that the majority of the studies that have assessed the comparability of CA formats reported results obtained from economic experiments conducted in hypothetical settings (with the exception of Chang et al., 2009; and Akaichi et al., 2013). Additionally, the experiments involving BWS as the preference elicitation method have been always implemented in hypothetical settings (Louviere et al. 2008; Scarpa et al. 2011; Lancsar et al. 2013). However, due to the skepticism surrounding the validity of values obtained from hypothetical CA experiments (Lusk and Schroeder 2004; Ding, et al. 2005; Alfnes et al., 2006; Lusk and Shogren, 2007; Dong et al., 2010), we conducted the CE, the RCA and the BWS in a non-hypothetical setting (further explanation is given in the section dedicated to the experimental design). In this study we also conducted a hypothetical CE to be used as the benchmark.

One of the most criticized issues of the experimental method is related to external validity. Whether, the behavior of experimental subjects reflects their actual purchase behavior. Therefore, to compare the external validity of the three CA formats, we have included a non-hypothetical holdout choice task in the experimental design in the CE, RCA and BWS. Finally, one of the main assumption underlying stated preference methods is that respondents know their preferences and these preferences are stable and coherent (Brown et al., 2008). According to Hoeffler and Ariely (1999), preferences' consistency or stability is positively correlated with choice experience and cognitive choice effort. For instance, in repeated choices, respondents are expected to be more precise and consistent in their decisions due to the learning effect (Brouwer et al., 2010). On the contrary, when they face hard choice tasks (e.g. too many choice sets or too many options per choice set), respondents are less precise and consistent in their choices compared with those facing an easy choice scenario (Brouwer et al., 2010). In this study, we compared the consistency of respondents' answers in CE, RCA and BWS to find out which of the three CA formats provide more consistent subjects' responses. To tackle this issue, one of the choice sets faced by respondents was repeated at the end of the choice experiment.

To sum up, our study stands out by assessing the comparability of non-hypothetical CE (NHCE), non-hypothetical RCA and non-hypothetical BWS in terms of estimated partworths, internal and external predictive power, estimated WTP, and participants' responses consistency, in the context that only the most preferred option is considered (RRCA and RBWS)[1]. This will allow us to assess the hypothesis whether the respondents behave similarly when state their most preferred option or when they state which of the option would choose, taking into consideration the BWS as substitute of RCA method.

This study is structured into five sections. In the next section, the experiment design and the experimental procedures are described. The econometric model used to estimate the partworths is outlined in section 3. The results are discussed in the fourth section and we finish by drawing some concluding remarks.

2. Experiment design

In this study, four treatments were carried out, hypothetical CE (HCE), NHCE, RRCA and RBWS. To assess the comparability of these CA formats a representative sample of 220 real consumers was recruited. Participants were randomly and equally assigned to the four treatments. Olive oil is the food product used in our experiment. The main attributes and attribute levels were first identified based on the literature review and the information collected from two focus groups of high and low experienced consumers of

[1] RRCA and RBWS stand for Recoded Ranking Conjoint Analysis and Recoded Best Worst Scaling, relatively. The data obtained in RCA and BWS were recoded as choice data (i.e only considering the option ranked first or the option chosen as the best option and recode it as the most preferred option).

olive oil. Four attributes of olive oil were considered. Three of them have three levels: type of olive oil (virgin extra, virgin, and olive oil)[2], origin (Andalucía, Catalonia, and rest of Spain) and price (2.20 €/liter, 3.50 €/liter, and 4.80 €/liter, which account for 85% of the price distribution in retail outlets). Brand is the fourth attribute and has two levels (Manufacturer label and Private label).

Given these attribute levels, a full factorial design of 54 (3^3*2) combinations (i.e. one-liter bottles of olive oil) was generated. Presenting respondents with 54 combinations, however, could place a high level of cognitive burden on respondents. To reduce the number of combinations that participants have to evaluate, we followed Street and Burgess (2007) and we generated an orthogonal fractional factorial design of 9 combinations. These 9 combinations were considered as the first option in each choice set. Since participants were provided with choice sets of 5 options each (plus a no-choice option), the other four options were obtained using the following generators (1000), (1111), (2121), and (2122) (Street and Burgess, 2007). This resulted in a 100% efficient main-effects design.

Participants, in each treatment, performed two choice tasks (i.e. main task and holdout task). In the main task and depending on the treatment (HCE, NHCE, RRCA or RBWS) were successively offered a total of 10 choice sets (i.e. first they received the 9 choice sets obtained in the efficient design. Then the fifth choice set was again given to participants to assess the consistency of their decisions). In each choice set, participants were asked to mark their most preferred option (or to rank all the options in RCA and BWS). The holdout task consists in a single-choice card of 10 options (including a no-choice option) which are different from the options provided to participants in the main task. In the holdout task, participants were required to choose the most preferred option of the 10 options included in the choice set. After finishing the two tasks, participants were asked to complete a short questionnaire about their socio-demographic and lexicographic characteristics as well as their attitudes toward olive oil.

At the beginning of the experiment, participants were informed that they would receive 15 Euros in cash at the end of the experiment. Additionally, participants in

[2] The three types of olive oil were defined according to the International Olive Council (IOC). Extra virgin olive oil: virgin olive oil which has a free acidity, expressed as oleic acid, of not more than 0.8 grams per 100 grams; Virgin olive oil**:** virgin olive oil which has a free acidity, expressed as oleic acid, of not more than 2 grams per 100 grams; Olive oil is the oil consisting of a blend of refined olive oil and virgin olive oils fit for consumption as they are. It has a free acidity, expressed as oleic acid, of not more than 1 gram per 100 grams. The three olive oils defined have other characteristics of which correspond to those fixed for this category in the IOC standard**.**

NHCE, RRCA and RBWS were informed that they would be participating in non-hypothetical tasks and, hence, it is in their best interest to reveal their actual preferences. We then explained them who the CA mechanism works.

Hypothetical (HCE) and non-hypothetical choice experiment (NHCE)

In HCE, we asked participants to assume that each choice set is a real shopping situation. Participants were informed, however, that they are not required to actually buy the chosen products and pay the corresponding price. In each choice set, participants were asked to indicate the option they prefer most bearing in mind their real purchase habits. They were also informed that if they did not like any one of the provided olive oil combinations, they simply can choose the no-choice option. The NHCE experiment was similar, but participants were informed that each choice set is a real shopping scenario. Therefore, participants could receive the option they had selected and pay its posted price. After finishing the main task, participants in both treatments were given a choice set of 10 options (i.e. holdout task) and were then asked to choose the option they prefer most.

After completing the two tasks and the survey, we asked for a volunteer among the participants to randomly draw a number between 1 and 2 to determine the binding task. If the binding task is the main task, participants in the HCE, receive the 15 Euros and the experiment finishes. In the NHCE, another volunteer is selected to randomly draw one of the 9 choice sets[3] to determine which of the choice set will be the binding one. Then, each participant obtains the option she/he has chosen in the binding choice set and receives 15 Euros minus the price indicated in that option. In case, the participant chose the no-choice option, she (he) receives the 15 Euros and do not buy any product. If the binding task is the holdout task, regardless the type of treatments (HCE or NHCE), each participant has to buy the chosen option and pays the corresponding price. If the chosen option is the no-choice option, the participant receives the 15 Euros and did not buy any product.

Non-hypothetical rank conjoint analysis (RRCA)

The same 10 choice sets were presented to each participant, who was asked to rank the options in each choice set from the most to the least preferred option. In case participants do not like any one of the presented alternatives, she (he) could choose the no-choice option. After completing all the choice sets in the main task, participants were

[3] The last choice set (the number 10) was the same as the fifth choice set. Therefore, to allow that all the choice sets have the same probability to be drawn we removed the tenth choice set.

given a choice set of 10 options (i.e. holdout task) and were then asked to choose the option they prefer most. After completing the main and the holdout task, a volunteer among participants was asked to randomly draw the binding task. If the main task was chosen as the binding task another volunteer was approached to draw the binding choice set. Following Lusk et al. (2008), to ensure that the ranking treatment is incentive compatible, participant had to purchase the binding product with a probability proportional to the rank she (he) assigned to each one of the options. Particularly, each participant who did not choose the no-choice option draws a number from 1 to 50 to select the biding product. If the number drawn is between 1 and 17, participant should purchase the most preferred option and pay its price. If the drawn number is between 18 and 30 the second most preferred option will be the biding product. If the drawn number is between 31 and 40, participant should purchase the third option in her (his) preference ranking and pay its corresponding price. If the number drawn is between 41 and 47, participant buys her (his) fourth most preferred option. Finally if the drawn number is between 48 and 50, participant has to buy the least preferred option. If the binding task was the holdout task, the procedure was similar to the one implemented in HCE and NHCE treatments.

Non-hypothetical best worst scaling (RBWS)

In BWS, each participant was asked to choose the most preferred or the best option, followed by the worst option of the four remaining options, followed by the second best option of the three remaining options, followed by the second worst option of the two remaining options. Hence, a complete ranking of the five options can be deduced (i.e. the option ranked first, second, third, fourth and fifth are the first best option, the second best option, the remaining option, the second worst option and the first worst option, respectively). Similar to the other three treatments and after finishing the main task, participants in NHBWS were given a choice set of 10 options (i.e. holdout task) and were then asked to choose the option they prefer most. Once participants finish the main and the holdout task, similar procedure to the one applied in the RCA was used to determine the binding task and the binding product.

3. Methodological approach

Based on the random utility theory (McFadden, 1973) and the Lancaster's (1966) theory (1966), the i^{th} individual's utility function U_{ijs} towards an option j from a choice set s can be decomposed into a deterministic component V_{ijs} and a stochastic component .

(1)

The deterministic component is commonly specified as linear in parameters and includes variables that represent the attributes of the alternative:

(2)

In Eq. (2) the attributes levels (the extra virgin olive oil (EVOO), the olive oil (OO), the Manufacturer Brand (BrManf), the Catalonian origin (CAT) and the "Rest of Spain" origin (RSp)) were effect coded (-1, 0, 1)[4], except for the price that was coded as a linear variable. The parameter "NoBuy" represents the no-choice option and has been coded as a dummy variable that takes the value 1 when the option was chosen by the participant; and 0, otherwise. To estimate the utility function (3), in HCE, NHCE, RRCA and RBWS, the random parameter logit model (RPL) was used to account for respondents' preferences heterogeneity. As shown by Train (2003), the probability of consumers *i* to choose the option *j* in the choice set *s* is as follows:

(3)

where is the density function of the coefficients . refers to the moments (the mean and standard deviation) of the parameter distributions and

(4)

It is worth noting that the estimation of RPL takes into account the unobserved effect of possible correlations between the attributes (Hensher et al., 2005). In fact, we assume that all the partworths of our empirical model are random and follow a normal distribution with mean and variance-covariance matrix , as they are not independently distributed.

Taking into account the between-subjects nature of our experimental design, it was necessary to test for the regularity of preferences across treatments. As in Lusk and Schroeder (2004), Caparros et al. (2008) and Lanscar et al. (2013), it is important to investigate whether differences in parameter estimates across samples are indeed due to the underlying preferences or to difference in variance. The null hypothesis of the test is the equality of preferences across treatments (i.e. , with is scale parameter). The test statistic is a likelihood-ratio type (, and it is distributed as a with K(M-1) degrees of freedom. is the log likelihood values

[4] The attribute levels virgin olive oil (VOO), private brand (BrPRV), and Andalucía (AND) were considered as the baseline for the attributes: type of olive oil, brand, and origin, respectively.

of the pooled data (e.g. HCE plus NHCE data), is the log likelihood values of the estimated model for each treatment, K is the number of restrictions, and M is the number of treatments (Louviere et al., 2000). If the hypothesis is rejected, comparing the estimated WTP for each treatment would be appropriate because the error variance is constant within each sample, and it will be simplified when calculating the marginal WTPs.

Willingness to pay

WTP estimates were calculated by dividing the estimated partworth associated with the attribute's level by the estimated partworth of the price attribute with a negative sign. To test the statistical of possible differences of the estimated WTP for each attribute across treatments, the non-parametric complete combinatorial test proposed by Poe et al. (2005) was used. This test first requires the generation of a distribution of 1000 WTP estimates using, for example, the parametric bootstrapping method proposed by Krinsky and Robb (1986). The complete combinatorial test is then applied to compare the 1000 bootstrapped WTP values in one treatment (HCE) with the 1000 bootstrapped WTP values in the other treatment (e.g. NHCE).

Consistency, internal and external validity

As aforementioned, to assess participants' responses consistency and their stability across treatments, the fifth choice set was repeated at the end of the main task[5]. To measure the consistency of participants' responses in each treatment, we calculated the proportion of participants who gave the same response in the fifth and the tenth choice set. The response is counted as a hit it is found to be the same in the fifth and the tenth choice sets. Then, the hit rate is calculated by dividing the total number of hits by the total number of participants in each treatment. To compare the hit rates across treatments, the Z-test was used.

To assess the internal validity of the estimated parameters we have used the estimated partworths to predict participant's response in the main task. The hit rate is calculated by comparing the predicted participants' decisions, using the maximum utility approach, to their real response in the fifth choice set in each treatment. Finally, as regard to the external validity, the estimated partworths from the main task are used to predict participants' responses in the holdout task. The predicted and the actual

[5] In line with Brower et al. (2010), respondents felt significantly more confident and certain about their choice at the end of the choice experiment than they were at the beginning. Therefore, repeating the fifth choice set at the end of the experiment instead the first one could increase the reliability of the test.

decision in the holdout task are compared to determine the hit rate. A Z-test was used to assess the difference between hit rates across treatments for both internal and external validity.

4. Results

Table 1 shows the number of times participants choose or ranked first each one of the options presented in the choice sets. As it can be observed, there exists a significant similarity between the choice experiment treatments (HCE vs. NHCE) as well as between the ranking treatments (RRCA vs. RBWS). In the HCE, respondents chose the first option as the best choice around 20% of the time, 14% of the time the second option, 17% of the time the third option, 17% of the time the fourth option, 16% of the time the fifth option and 16% of the time the no-choice option. These percentages are statistically similar to those obtained in NHCE (i.e. 16%, 15%, 17%, 18%, 16% and 18% for the first to the fifth option plus the no-choice option, respectively). A similar result was found when comparing the ranking treatments. In the RRCA, respondents chose the first option 23% of the time, the second option 15% of the time, the third option 21% of the time, the fourth option 17% of the time, the fifth option 17% of the time and the no-choice option the 6% of the time as the most preferred option; these percentages are not statistically different from those found in the RBWS. However, Table 1 also shows that there are significant differences in participant choices between CE treatments (HCE and NHCE) and ranking treatments recoded as traditional choice (RRCA and RBWS).

The second way of comparing results from the four treatments is to analyze the means and the standard deviations of the estimated partworths. In the Table 2 we report the estimates corresponding to the four treatments where the dependent variable has been coded in a similar way. Particularly, in HCE and NHCE, the dependent variable takes the value of 1 when the option is the chosen one and 0 otherwise; while in RRCA and RBWS the dependent variable is coded as 1 when the option is ranked first and 0 otherwise. Results reveal that the no-choice option parameter is negative and significant across treatments, indicating that the majority of the respondents opted for choosing or ranking the real options instead of selecting the no-choice option. Additionally, results in Table 2 shows that all random parameters have the expected sign and are statistically significant as well as their standards deviations. However, the lower magnitude of the standards deviations, in most of cases, suggests the non-existence of significant high preference heterogeneity between respondents.

Table 1. Comparison of choices across the treatments

	Treatments			
Options	% of time participants choosing each option as the best choice			
	HCE	**NHCE**	**RRCA**	**RBWS**
1	19.79	16.16	23.23	23.63
2	13.73	14.54	15.35	13.93
3	16.76	17.37	21.21	17.17
4	17.17	17.97	16.96	17.97
5	16.16	16.36	17.17	20.20
No option	16.36	17.57	6.06	7.07
Test for equality proportions	*Chi-square: 2.30* *p-value: 0.80*		*Chi-square: 4.20* *p-value: 0.52*	
	Chi-square: 62.02 *p-value: 0.00*			

Interestingly, the results reveal that participants' preferences for the attributes of olive oil in the four CA formats are similar in terms of sign and significance although their estimated values are different. In fact, our findings (Table 2) show that the extra virgin is most preferred olive oil for consumers, while the olive oil is the least preferred. We also found that the local olive oil (Catalonian olive oil) is preferred over the olive oil from the other locations, while olive oil produced in Andalusia (i.e. South of Spain) is preferred over olive oil produced in the rest of Spain. This result is consistent with the findings of Jiménez-Guerrero et al. (2012) who found that the origin of olive oil is key attribute for Spanish consumers. Furthermore, we found that consumers seem to prefer the manufacturer brand over the private label. Finally, our results show that price is the main obstacle for buying olive oil.

A likelihood ratio test has been used to test for significant differences among the estimated partworths across the 4 treatments. The results of likelihood ratio test are displayed in Table 3. Results show that we could reject the null hypothesis of equality of the estimates obtained in HCE and NHCE (LR= 49.05; $p < 0.005$). Furthermore, we compare the results from the three non-hypothetical treatments taking into account only the most preferred option. In this case, the null hypothesis is clearly rejected (LR = 163.89; $p < 0.005$). The final tests compare the estimates only from the two ranking treatments RRCA and RBWS. The null is also rejected (LR = 66.64; $p < 0.005$ and LR = 66.87; $p < 0.005$, respectively). Therefore, the results pointed out that the differences between the estimates across all treatments are statistically significant. Consequently, are these discrepancies in the partworths' values is going to affect the comparability of the four conjoint analysis formats in terms of internal and external validity, as well as participants' WTP.

Table 2. RPL and RO-RPL models estimates for elicitation methods

	Only choice or first Rank data in ranking treatments			
Traitment	**HHCE**	**NHCE**	**RRCA**	**RBWS**
Models	**RPL**	**RPL**	**RPL**	**RPL**
	Random parameter estimates			
NoBuy	-4.82***	-3.34***	-5.05***	-5.01***
(SE)	(0.47)	(0.35)	(0.46)	(0.46)
EVOO	1.55***	0.92***	0.79***	0.98***
	(0.34)	(0.21)	(0.17)	(0.20)
VOO[a]	-0.20	0.09	0.08	-0.11
	(----)	(----)	(----)	(----)
OO	-1.35***	-1.02***	-0.88***	-0.87***
	(0.27)	(0.24)	(0.17)	(0.23)
BrManf	0.34***	0.21**	0.15**	0.18**
	(0.09)	(0.08)	(0.07)	(0.08)
BrPrv[a]	-0.34	-0.21	-0.15	-0.18
	(----)	(----)	(----)	(----)
AND[a]	-0.65	-0.33	0.11	-0.39
	(----)	(----)	(----)	(----)
CAT	1.30***	0.91***	0.55***	0.94***
	(0.19)	(0.19)	(0.13)	(0.16)
RSp	-0.65***	-0.57***	-0.66***	-0.54***
	(0.17)	(0.17)	(0.13)	(0.13)
Price	-1.60***	-1.21***	-0.69***	-1.08***
	(0.16)	(0.11)	(0.11)	(0.11)
	Standards deviations of random parameters			
EVOO	2.49***	1.77***	1.46***	2.19***
	(0.31)	(0.22)	(0.17)	(0.25)
OO	2.94***	1.71***	1.36***	2.63***
	(0.43)	(0.21)	(0.19)	(0.32)
BrManf	0.14	0.28**	0.22**	0.41**
	(0.13)	(0.11)	(0.08)	(0.13)
CAT	2.14***	1.32***	0.68***	1.14***
	(0.28)	(0.18)	(0.13)	(0.18)
RSp	0.77***	0.99***	0.50**	0.63***
	(0.19)	(0.18)	(0.16)	(0.16)
Price	1.62***	0.49***	1.15***	1.12***
	(0.20)	(0.07)	(0.11)	(0.13)
Number of observations	2970	2970	2970	2970
Log-likelihood	-523.21	-591.45	-619.10	-575.72

[a] base line; (***) (**) (*) Statistically significant at 1%, 5%, and 10% level.

The results of responses' consistency, internal and external validity analysis are displayed in Table 4. The results reveal that, in general terms, the consistency of participants' responses is relatively high. The hit rate ranges from 76.36% to 87.27% when only the most preferred option is considered (i.e. HCE, NHCE, RRCA, and RBWS). Results from the one tailed Z-test for consistency, using the hit rates show that there are no statistical differences (at the 5% level of significance) between treatments. Consistent with Brouwer et al. (2010), the high consistency or stability of participants' responses could be due to the learning effect related to repeated choice task. Additionally, we found that the consistency of participants' responses in HCE and NHCE is statistically similar. This similarity maybe the result effort made by the experimenter to incentivize participants to reveal their real preferences independently of whether they are taking part in a hypothetical or a non-hypothetical CE.

Table 3. Results from preference regularity tests across the treatments

Test for preference regularity	Number of observations	Log Likelihood	Likelihod Ratio (LR)	Degrees of freedom	p-value
All treatments	5940	-1139.11			
HCE	2970	-523.21			
NHCE	2970	-591.45			
H_0:test of equality between hypothetical and non-hypothetical CE			49.05	12	P<0.005
All treatments	11880	-2413.84			
HCE	2970	-523.21			
NHCE	2970	-591.45			
RRCA	2970	- 619.10			
RBWS	2970	- 575.72			
H_0:test of equality between hypothetical and non-hypothetical first choice option			208.66	72	P<0.005
All treatments	8910	-1868.23			
NHCE	2970	-591.45			
RRCA	2970	- 619.10			
RBWS	2970	- 575.72			
H_0:test of equality between non-hypothetical first choice option			163.89	36	P<0.005
All treatments	5940	-1228.15			
RRCA	2970	- 619.10			
RBWS	2970	- 575.72			
H_0:test of equality between non-hypothetical NHRCA and NHBWS			66.64	12	P<0.005

As regard to the internal validity (central part of table 4), results are quite similar across treatments. The hit rate ranges from 62% to 73% indicating that, in the main task, the estimated partworths correctly predicted 62% to 73% of participants' responses. Results from the one tailed Z-test show that no significant differences have been found between the four CA formats in terms of internal validity.

Regarding the external validity, the estimated parameters in the hypothetical CE correctly predicted only 23.63% of participants' responses in the holdout task, which is significantly lower than the predictive power of the estimates obtained in the non-hypothetical CA formats. Incentivizing participants to truthfully reveal their preferences seems to enhance the predictive power of the estimated models (Chang et al. 2009; Lusk et al. 2008). Consistent with the findings of Akaichi et al. (2013), we found that the external validity of the estimates in the non-hypothetical CA formats is similar when the responses in RCA and BWS are coded as choice data taking into account only the option ranked first. Although, the external validity in RBWS is higher than NHCE and RRCA (i.e. the external validity is 52.63%, 38.18%, and 43.63% in RBWS, RRCA, and NHCE, respectively), however this gain in predictive power is not statistically significant at 5%.

Since one of the main reasons of using choice experiments is to estimate consumers' willingness to pay for specific food attributes. The results of the comparability of the four CA formats in terms of WTP are displayed in Table 5. As it can be observed, consumers are willing to pay a price premium ranging from 0.76 Euro and 1.14 Euro for the extra virgin olive oil; however, they are willing to pay a lower price for the olive oil with respect to virgin olive oil (baseline level). In relation to the origin, respondents were willing to pay a price premium (varies between 0.75 and 0.86 euro across treatments) for the olive oil from Catalonia. Finally, on average, consumers are willing to pay an average premium of about 0.2 Euro for the manufacturer brand with respect to the private brand.

Results displayed in Table 6 show a strong similarity of the estimated WTP values across the CA formats. Therefore, the statistical differences we have found among partworths estimates are indeed underlying differences in the error variances due to difference cognitive effort spent by respondents in the different CA formats. Carlsson and Martinsson (2001) and Lusk and Schroeder (2004), also arrived to a similar conclusion when they compared hypothetical and non-hypothetical CE. Two potential explanations could be given: 1) the inclusion of the no-choice option freed participants from being forced to choose a real option for which they have to pay the corresponding price (Hensher 2010); and 2) the effect of a well-script presentation explaining the objectives and the characteristics of each treatment at the beginning of every session has contributed to reduce the hypothetical bias (Hensher, 2010; Murphy et al., 2005). In relation to the non-hypothetical treatments, Akaichi et al. (2013) arrived to a similar result when they compared the NHCE and the RRCA.

Table 4. Consistency, internal and external validity tests across treatments

	Consistency				Internal validity			External validity		
Treatments	**Nº of choices**	**Nº of correct Predictions**	**Hit rate (%)**	***p-value***	**Nº correct of predictions**	**Hit rate (%)**	***p-value***	**Nº correct of predictions**	**Hit rate (%)**	***p-value***
HHCE vs	55	48	87.27	0.10	35	63.63	0.15	13	23.63	0.01
NHCE	55	43	78.18		40	72.72		24	43.63	
HHCE vs	55	48	87.27	0.06	35	63.63	0.15	13	23.63	0.04
RRCA	55	42	76.36		40	72.72		21	38.18	
HHCE vs	55	48	87.27	0.06	35	63.63	0.42	13	23.63	0.00
RBWS	55	42	76.36		34	61.81		29	52.72	
NHCE vs	55	43	78.18	0.41	40	72.72	0.50	24	43.63	0.28
RRCA	55	42	76.36		40	72.72		21	38.18	
NHCE vs	55	43	78.18	0.41	40	72.72	0.11	24	43.63	0.17
RBWS	55	42	76.36		34	61.81		29	52.72	
RRCA vs	55	42	76.36	0.50	40	72.72	0.11	21	38.18	0.06
RBWS	55	42	76.36		34	61.81		29	52.72	

Table 5. Estimated Willingness to Pay for each attribute level

Traitment	HHCE	NHCE	RRCA	RBWS
EVOO	0.97*	0.76*	1.13*	0.90*
	[0.55; 1.38]	[0.38; 1.14]	[0.51; 1.76]	[0.52; 1.28]
OO	-0.84*	-0.84*	-1.26*	-0.80*
	[-1.15; -0.53]	[-1.24; -0.43]	[-1.86; -0.66]	[-1.23; -0.37]
BrManf	0.21*	0.17*	0.22	0.17*
	[0.10; 0.33]	[0.02; 0.32]	[-0.001; 0.45]	[0.01; 0.33]
CAT	0.81*	0.75*	0.79*	0.86*
	[0.54; 1.08]	[0.43; 1.07]	[0.36; 1.21]	[0.56; 1.16]
RSp	-0.40*	-0.47*	-0.95*	-0.50*
	[-0.63; -0.18]	[-0.75; -0.19]	[-1.41; -0.49]	[-0.74; -0.26]

Values in brackets correspond to Confidence Intervals;
An * indicates that the value is statistically different from zero at the 5% siginicance level.

To sum up, this is study is the first attempt to compare the four most used CA formats (i.e. HCE, NHCE, RRCA and RBWS). Our results suggest that, the four CA formats provide similar results in terms of the sign and the significance of estimated partworths as well as the estimated WTP values. Therefore, if the estimation of consumers' WTP is the main objective of using CA, then the use of any one of the four CA format is appropriate. Nonetheless, the use of HCE might be preferred due to the simplicity and lower cost of its implementation. Furthermore, if practitioners are interested in assessing consumers' preferences and the use of the estimated partworths for prediction sakes (e.g. to predict the change in consumers' choices when the price changes), then non-hypothetical BWS maybe should be used due to the gain of predictive power that it add in respect to NHCE and RRCA, although this difference was not statistically significant at 5%.

Table 6. Hypothesis test of equality WTPs across the treatments

Hypothesis	**EVOO**	**OO**	**BrManf**	**CAT**	**RSp**
	p-value of Complete combinatorial test				
HCE vs NHCE	0.47	0.43	0.48	0.48	0.45
HCE vs RRCA	0.48	0.47	0.42	0.48	0.49
HCE vs RBWS	0.43	0.43	0.39	0.48	0.46
NHCE vs RRCA	0.44	0.45	0.42	0.45	0.44
NHCE vs RBWS	0.40	0.37	0.38	0.45	0.41
RRCA vs RBWS	0.45	0.41	0.44	0.49	0.46

5. Conclusions

Focusing on a market good, the olive oil, this study, aimed to compare HCE, NHCE, RRCA and RBWS in terms of consistency, estimated partworths, internal and the external validity and estimated WTP. This study expands the previous works in three ways: 1) it includes an additional ranking CA format (i.e. BWS) which has been proved to be easily understood by consumers when they are asked to rank several combinations of attributes; 2) the BWS is conducted in a non-hypothetical setting adopting the approach used by Lusk et al. (2008) for RCA; and 3) in this study we have compared the consistency among respondents' decisions through the three response formats used..

Results from this study suggest a number of points. First, the estimated partworths from the four CA formats are similar in terms of sign and significance although they have different values. For instance, all the CA format showed that responded preferred the local extra virgin oil. Additionally, the results show that there are not significant differences across the treatments in terms of estimated WTP. This could be used as an argument in favor of the use of hypothetical choice experiment especially, when the main reason for using the CE is to assess consumers' WTP.

Second, using RCA or BWS could be advantageous if the practitioner is interested in determining not only the most preferred option but also consumers' preferences for all the combinations of attributes included in a choice set. In fact, carrying out CE, RCA and BWS in a non-hypothetical context and considering only the most preferred option in the estimation of partworths, they provide similar results in terms of partworths and WTP. However, the use of RCA and BWS might be preferred since they provide practitioner with information all the options included in a choice set.

Finally, our results showed that the respondents behave similarly when they asked to state their preferred options or to choose their most preferred options, taking into consideration two ranking preference methods RCA and BWS. However, conducting BWS in non-hypothetical context seems to outperform RCA and could classify it as competitive substitute of RCA method. As aforementioned, this superiority might be a result of the fact that BWS better fits natural human skills at identifying extreme values that decreases the cognitive burden on participants which in turn could increase the predictive power of the estimates.

It noteworthy that more research work is, however, needed to see to what extent our results can be generalized. For instance, we think it important to know how our findings will be affected if large choice sets are used. Furthermore, additional empirical applications with alternative products should be carried out to find out whether our findings are sensitive to the type product (e.g. food vs. non-food product or durable vs. fresh food) used in the experiment. Finally, all the previous studies compared the CA formats make their comparison based only on the most preferred option; therefore, it is important to assess the comparability of the three CA formats (CE, RCA and BWS) taking into consideration the most preferred option as well as the respondents' full ranking preferences.

References

Akaichi, F., R. M. Nayga, Jr. and J. Gil. "Are Results from Non-Hypothetical Choice-Based Conjoint Analysis and Non-Hypothetical Recoded-Ranking Conjoint Analysis Similar?" *American Journal of Agricultural Economics* Vol. 95, (2013) pp. 949-963.

Alfnes, F., Guttormsen, A.G., Steine, G., and Kolstad. K. "Consumers' willingness to Pay for the Color of Salmon: A Choice Experiment with Real Economic Incentives." *American Journal of Agricultural Economics* Vol. 88, (2006) pp. 1050-1061.

Ben-Akiva, M., Morikawa, T., and Shiroishi, F. "Analysis of the reliability of preferences ranking data" *Journal of Business Research* Vol. 24, (1992) pp. 149-164.

Boyle, K.J., Holmes, T.P., Teisl, M.F., and Roe., B. "A Comparison of Conjoint Analysis Response Formats." *American Journal of Agricultural Economics* Vol. 83, (2001) pp. 441–54.

Brouwer, R., Dekker, T., Rolfe, J., and Windle, J. "Choice certainty and consistency in repeated choice experiments" *Environmental Resource Economics* Vol. 46, (2010) pp. 93-109.

Brown, T.C., Kingsley, D., Peterson, G.L., Flores, N.E., Clarke, A., and Birjulin, A. "Reliability of individual valuations of public and private goods: choice consistency, response time, and preference refinement" *Journal of Public Economic* Vol. 92, (2008) pp. 1595-1606.

Campbell, D., and Lorimer, V.S. "Accommodating attribute processing strategies in stated choice analysis: do respondents do what they say they do?" presented at European Association of Environmental and Resources Economist Annual Conference, Amsterdam, June 2009.

Caparrós A., Oviedo, J.L., and Campos., P. "Would you Choose your Preferred option? Comparing Choice and Recoded Ranking Experiments." *American Journal of Agricultural Economics* Vol. 90, (2008) pp. 843-855.

Carlsson, F., and Martisson, P. "Do hypothetical and actual marginal willingness to pay differ in choice experiments?" *Journal of Environment Economics and Management* Vol. 41, (2001) pp. 179-192.

Chang, J.B., Lusk, J.L., and Norwood, B. “How closely do hypothetical surveys and laboratory experiments predict field behavior?” *American Journal of Agriculture Economics* Vol. 96, (2009) pp. 518-534.

Ding, M. “An incentive-aligned mechanism for conjoint analysis” *journal of Marketing Research* Vol. 44, (2007) pp. 214-223.

Ding, M., Grewal, R., and Liechty, J. “Incentive-Aligned Conjoint Analysis.” *Journal of Marketing Research* Vol. 42, (2005) pp. 67–83.

Dong, S., Ding, M., and Huber, J. “A simple mechanism to incentive-align conjoint experiment” *International Journal of Research in Marketing*, Vol. 27, (2010) pp. 25-32.

Flynn, T., and Marley, A.J. (2012) “Best worst scaling: Theory and methods” Sydney, CenSoC Working Paper Nº 12-002.

Helson, H.. *Adaptation-Level Theory* (New York: Harper & Row, 1964).

Hensher, D.A. “Hypothetical bias, choice experiment and willingness to pay” *Transport Reseach Part B* Vol. 44, (2010) 735-752.

Hensher, D.A., Rose, J.F., Greene, W.H. *Applied Choice Analysis: A Primer.* (Cambridge University Press, Cambridge 2005).

Hoeffler, S., and Ariely, D. “Constructing stable preferences: a look into dimensions of experience and their impact on preferences stability” *Journal of Consumer Psycologic* Vol. 8, (1999) pp. 113-126.

Jiménez-Guerrero, J.F., Gásquez-Abad, J.C., Mondéjar-jiménez, J.A. and Huertas-García, R. “consumer preferences for olive oil attributes: a review of the empirical literature using a conjoint approach” chapter 12 in “*olive oil constituents, quality, health properties and bioconversions*”, edited by Boskou Dimitrios, (Publisher: In Tech. 2012)

Krinsky, I., and Robb, A.L. “On Approximating the Statistical Properties of elasticities.” *Review of Economics and Statistics* Vol. 64, (1986) pp. 715–19.

Lancaster, K. J. “A new approach to consumer theory”, *The journal of political economy*, Vol. 74, (1966) pp.132-157.

Lancsar, E., Louviere, J., Donaldson, C., Currie, G., and Burgess, L. “Best worst discrete choice experiment in health: Methods and an application” *Social Science and Medicine* Vol. 76, (2013) pp. 74-82.

Louviere, J., Street, A., Burgess, L., Wasi, N., Islam, T., and Marley, A. “ Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information" Sydney, CenSoC Working Paper Nº 04-005, (2004).

Louviere, J.J. and Street, D. Stated-preference Methods in David A Hensher & Kenneth J Button (eds), *Handbook of Transport Modelling*, (Pergamon Press, Amsterdam, Netherlands, 2000).

Louviere, J.J., Street, D., Burgess, L., Wasi, N., Towhidul, I., Anthony, A.J.M. "Modelling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information" *Journal of Choice Modelling*, Vol. 1, (2008) pp. 128-163.

Lusk, J.L., and Schroeder., T.C. "Are Choice Experiments Incentive Compatible? A Test with Quality Differentiated Beef Steaks." *American Journal of Agricultural Economics* 86, (2004) pp. 467–82.

Lusk, J.L., and Shogren, J.F. *Experimental auction: Methods and applications in economic and marketing research* (Cambridge: Cambridge University Press, 2007).

Lusk, J.L., Fields, D., and Prevatt. W. "An Incentive Compatible Conjoint Ranking Mechanism." *American Journal of Agricultural Economics* Vol. 90, (2008) pp. 487–98.

McFadden, D. *Conditional logit analysis of qualitative choice behavior*, En Zarembka, P. (ed.) Frontiers in econometrics. (Academic press, Nueva York, 1973).

Murphy, J.J., Geoffrey, J.A., Stevens, T.H., and Weatherhead, D. "Meta-Analysis of Hypothetical Bias in Stated Preference Valuation." *Journal of Environmental and Resource Economics* Vol. 30, (2005) pp. 313–325.

Pignone, M.P., Brenner, A.T., Hawley, S., Sheridan, S.L., Lewis, C.L., Jonas, D.E., and Howard., K. "Conjoint Analysis Versus Rating and Ranking for Values Elicitation and Clarification in Colorectal Cancer Screening." *Journal of General Internal Medicine* (published on line August 26th 2011).

Poe, G.L., Giraud, K.L., and Loomis, J.B. "Computational Methods for Measuring the Difference of Empirical Distributions." *American Journal of Agricultural Economics* Vol. 87, (2005) pp. 353-365.

Scarpa, R., Nataro, S., Louviere., J., and Raffaelli, R. "Exploring scale effects of best/worst Rank ordered choice data to estimate benefits of tourism in alpine grazing commons" *American Journal of Agricultural Economics*, Vol. 93, (2011) pp. 813-828.

Street, D., and Burgess, L.B. *The Construction of Optimal Stated Choice Experiments: Theory and Methods*, 1st ed, John Wiley and Sons, (Hoboken, New Jersey, 2007).

Train, K., *Discrete Choice Methods with Simulation* (Cambridge University Press, Cambridge 2003).