Economics Bulletin

Volume 30, Issue 2

Testing the cognitive burden of two choice modeling valuation variants. The between and within sample approaches

Pierre-Alexandre Mahieu
University of Rouen

Pere Riera
Autonomous University of Barcelona

Raul Brey
Pablo de Olavide University

Abstract

Scores are commonly used in environmental valuation exercises. The two main procedures when testing for score differences are the within sample and the between sample approaches. Their conclusions do not always coincide. With a case study involving scores on difficulty of responding to two choice modeling variants –contingent ranking and contingent grouping–, the paper shows the strength of the within sample approach when relying on the coherent arbitrariness principle. Results suggest that the grouping is significantly less difficult to complete than the ranking task. The validity of these results is enhanced by the fact that they are independent of the exercise order, which is tested by randomizing the sequence order in which respondents face the two methods.

The authors thank Emmanuel Flachaire for his helpful comments. This research was partly supported by the EU project VULCAN (EVK2-CT-2000-00094).

Citation: Pierre-Alexandre Mahieu and Pere Riera and Raul Brey, (2010) "Testing the cognitive burden of two choice modeling valuation variants. The between and within sample approaches", *Economics Bulletin*, Vol. 30 no.2 pp. 1384-1391.

Submitted: Apr 02 2010. **Published:** May 16, 2010.

1. Introduction

Environmental valuation surveys may include questions with scores. For instance, it is not uncommon to inquire in the following manner: "from 1 to 10, how sure do you feel about your answer?", or "in a 1 to 5 scale, how difficult was it to answer", or "could you rate the following alternatives?" One issue regarding these questions is that people might attach different meanings to the scale, as has been pointed out by several authors. Mitchell and Carson (1989) stress that meaning is subjective and contextual; Mackenzie (1993) states that "some respondents use the entire scale specified by the researcher, while others confine their ratings to different portions of that scale" (page 593). It has also been shown that scores are sensitive to cues such as the amount of the numerical points composing the scale (Schwarz *et al.*, 1991) or the graphical representation of the scale, e.g. horizontal or vertical (Friedman and Friedman, 1994).

A way to mitigate some of these problems may be through the use of reference points. Following a psychology perspective, Ariely *et al*. (2003) find in a valuation survey that individuals follow some *coherence arbitrariness*. That is to say that in their first answer respondents might be somehow arbitrary, but the subsequent answers are coherent relative to the first one. For instance, even if when estimating the value of a good wine, respondents might state a value based on some heuristics, rather than in their true WTP, if asked to value a regular wine afterwards, the stated amount tends to be coherently lower than the first one. The same logic applies to scores. If in a question involving a 1 to 10 scale, from "completely uncertain" to "completely certain", a respondent states a 7, a subsequent similar question for a different issue where the respondent is less certain might be answered by a 6. However, if presented in the reverse order, the first question might be answered by a 7 and the second by an 8, depending on how the respondent interprets the scale. Intuitively, the comparison of scores between questions seems to be more informative than analyzing the scores of each question separately.

Scores can also be used to compare the difficulty of different tasks, or valuation methods. This is often implemented in a split sample manner, where part of the sample receives a questionnaire version with a given task or valuation method, and the other part an alternative task or valuation method (for recent examples, see Caparros *et al.*, 2008; Whynes *et al.*, 2007; Yadav *et al.*, 2007). Comparisons are generally based on the mean score, allowing for *between sample* comparisons. An alternative is to assign two exercises to each individual and compare the scores given by the same individuals. This constitutes a *within sample* comparison.

The within sample approach overcomes some problems. When two subsamples are compared, individual differences between the two subsamples, rather than the difficulty of the valuation task, might be responsible for differences in scores. On the other hand, within sample approaches suffer from some drawbacks. The most common is probably the so-called "order effect", which has been already demonstrated in valuation surveys (Bateman and Langford, 1997). It implies that the first rating might influence the second one, thus suggesting that scoring is not independent from the question order. This can be tested by randomizing the order in which the different tasks are presented to respondents and applying a within sample test to each subsample separately.

The between sample and within sample with randomized succession order tests are applied to cognitive burden scores for two variants of choice modelling techniques in a survey on climate change effects over shrublands in Spain. The two valuation variants are the contingent ranking, consisting in ranking different alternatives given to respondents, and contingent grouping, where respondents group alternatives as better or worse than the business-as-usual situation. Both are explained at the beginning of section 3. Furthermore, the article discusses

how to draw the most likely conclusions based on these tests and according to the "coherent arbitrariness" principle (Ariely *et al.*, 2003) which suggests that people's valuation might be arbitrary in the first score but nevertheless coherent with it in the subsequent ones. In that regard, the paper highlights the advantages of the within sample approach and shows that its validity can be reinforced by randomizing the valuation tasks order.

This paper builds on Brey *et al*. (2007), but differs in the theoretical framework (the coherent arbitrariness principle) and the tests used, which lead to stronger results. Section 2 explains in more details the coherent arbitrariness principle. Section 3 introduces the tests. Section 4 describes the valuation case study. The main results and their discussion are presented in sections 5 and 6 respectively, while conclusions and further research constitute section 7.

2. Coherent arbitrariness

By means of an experiment and a review of valuation literature, Ariely *et al*. (2003) show that although a first stated WTP might be arbitrary within a range, a second stated WTP tends to be coherent with the first one. In the same manner, when faced with a series of bid amounts, an individual may not know whether her WTP is superior or inferior to the first bid, and give a heuristic response. But once she responds, the subsequent questions will be answered as if her preferences had been well formed. The authors argue that preferences would be initially "malleable", as indicated by the anchoring effect, but would become "imprinted" once a first amount is stated. The malleability of preferences has been modelled by Flachaire and Hollard (Flachaire and Hollard, 2007) through the range model, implying that respondents may consider a range of possible WTP, rather than a point. The existence of this range has been supported by recent empirical studies (Hanley *et al.*, 2009).

The coherent arbitrariness is not limited to monetary valuation, and may occur when no money is involved as shown by Ariely *et al* (2003). In one of the surveys, the participants were exposed to two stimuli: a sample of an unpleasant liquid (half *gatorade*, half vinegar) and an aversive sound. After experiencing them, they were to make a hypothetical choice, which was to drink another sample of the beverage or listen to the sound again. Then, in a follow up exercise, they were to state whether they would be willing to endure the sound for 10 seconds, 20 seconds, 30 seconds, etc., up to eight minutes, to avoid drinking a given quantity of liquid. Results were interpreted as demonstrating a coherent arbitrariness. This phenomenon may also apply in other contexts, such as when comparing the difficulty of two different tasks, as will be discussed below.

3. Tests

The application presented here uses two variants of the choice modelling valuation methods. One is the Contingent Ranking (CR) (Louviere *et al.*, 2000) and the other is the Contingent Grouping (CG) (Brey *et al.*, 2005). In each method application –hereafter also referred to as exercise –, a choice set with four alternatives is presented to respondents, the business-as-usual (BAU) situation being one of them. The choice task differs between exercises. For CR, the alternatives are to be ranked by order of preference, whereas for CG the non-BAU alternatives are to be grouped as better or worse than BAU –i.e., the respondent points out which alternatives she agrees with and which she discards, compared to keeping BAU.

Each respondent is faced with both exercises. Half of the sample sees the CR exercise first, followed by CG, and the other half sees the two exercises in the reverse order. This gives rise to two groups of respondents, the one from participants facing CG in the first round and CR in the second (the CGCR group), and the other from those confronted with CR first followed by

CG (the CRCG group). After completing each choice variant, respondents are asked to grade the difficulty encountered on a scale ranging from 1 ("very easy") to 7 ("very difficult"). This paper only focuses on the use of the difficulty scores, leaving out other differences between valuation methods. Two sets of tests are undertaken to check whether methods differ in choice task difficulty.

(i) First round means comparison

Difficulty scores from the two exercises when they appear first (hereafter "first round" scores) are compared. This constitutes a between sample comparison. The null and alternative hypotheses can be written as

$$H_0: \mu_R - \mu_G = 0$$
$$H_1: \mu_R - \mu_G \neq 0,$$

where μ_R and μ_G respectively denote the mean of the distribution of CR and CG difficulty scores in the first round. The rejection of the null hypothesis would suggest that one task is perceived as more difficult than the other.

(ii) Paired comparison with randomized sequential order

Scores are compared separately in each group (CGCR and CRCG) with paired-comparison tests, the difference score for each individual being the score given to the ranking exercise minus the score to the grouping one. The null and alternative hypotheses of this within sample comparison are in both cases expressed as

$$H_0: \mu_{(R-G)} = 0$$
$$H_1: \mu_{(R-G)} \neq 0,$$

where $\mu_{(R-G)}$ represents the mean of the differences between the ranking and grouping scores from each individual. The rejection of the null hypothesis would suggest that one task is perceived as more difficult than the other.

4. Application

A survey was administered in the region of Catalonia, Spain, to 354 individuals in 2004. Several investment programs aiming at mitigating the effects of climate change on shrublands were presented to the participants. Each investment program altered the BAU situation according to four attributes: density of the shrub vegetation expressed in percentage of plant cover (40%, 60%, and 80%), level of erosion expressed in percentage of shrubland soils subject to severe erosion (16%, 24%, and 32%), average percentage of shrubland annually affected by fires (3%, 4%, and 5%), and an annual payment to fund the program (5€, 15€, and 30€).

The different attributes and levels gave rise to 81 (3^4) possible combinations or alternatives, which were randomly distributed to produce 27 choice sets of three non-BAU alternatives. The BAU option was then added to all choice sets, with specific values: 40% for density, 32% for erosion, 5% for fires, and a payment of 0€ (the expected situation in 50 years with no investment). Thus, each choice set was finally composed of 4 alternatives.

Interviews were conducted face-to-face at people's homes, using laptop computers. Respondents read the screens and listened to a recorded voice. They typed in by themselves or dictate to the interviewer, at their discretion. No significant problems were detected in the survey application.

5. Results

(i) *First round means comparison*

A *t*-test is used to compare the means of the first round scores. Results are shown in Table I. Assuming a normal approximation for the *t*-statistic, the null hypothesis cannot be rejected at 5% level (*p*-value = 0.988). This suggests that there is no difference in difficulty scores between the two valuation methods.

Table I Unpaired *t*-test for the first round rating question

	Sample Size	Mean	Std. Error	P-value
Ranking scores	180	2.139	0.135	0.988
Grouping scores	174	2.144	0.144	

(ii) *Paired comparison with randomized sequential order*

A paired *t*-test is used to compare differences in scores within each group or subsample. For both groups, scores assigned to the CR exercise are significantly higher at 5% level than those attributed to CG, as shown in Table II. The null hypothesis is rejected, suggesting that ranking is found more difficult to answer than grouping.

Table II Paired *t*-test for each group

	Sample Size	Mean difference	Std. Error	P-value
CGCR group	180	0.253	0.071	0.000
CRCG group	174	0.256	0.081	0.001

A Mann Whitney test is used to compare the medians. This non parametric test which is also called Wilcoxon rank sum test can either be applied on a single sample, as for the within sample comparison, or on two samples, as for the between comparison. The same conclusions are found as for the mean comparison. The null hypothesis is rejected for the within sample comparison (p-value = 0.000 in each group), and fails to be rejected for the between sample comparison (p-value = 0.638).

It is also tested whether, in each group, there are more people stating a higher score for the ranking task than for the grouping task. Results of a *t*-test show that the proportion of participants stating a higher score for the ranking task is statistically larger in each group (see Table III). Again, this suggests that ranking is a more difficult task than grouping.

Table III Comparison of proportions of higher scores for each group

	Higher score to the grouping task	Higher score to the ranking task	P-value
CGCR group	0.080	0.236	0.000
CRCG group	0.078	0.211	0.000

6. Discussion

The results from the previous tests do not point in the same direction. The within sample comparison suggests that ranking is a more difficult task than grouping, whereas the between sample comparison suggests that ranking is as difficult as grouping. One of these approaches may imply a misleading conclusion.

The between sample approach relies on the assumption that individuals perceive and use the scale in the same fashion. This might however not be the case (Hensher *et al.*, 2005). Some people may be tempted to discard extreme points of the scale, as suggested by Mackenzie (1993). Moreover, the meaning of the endpoint label or the numerical points is likely to differ among individuals. For instance, a score of 2 on a 7 point scale might mean "very easy" to one participant and "somewhat easy" to another one. Besides, respondents might be unsure on the meaning to be associated to each numerical point. The vagueness of the endpoint labels may be partly responsible for it. Individuals may then rely on some cues to interpret the scale, like the value taken by the numerical points. Schwarz (1991) shows that a scale ranging from 1 to 10 does not yield the equivalent results to using a scale from -5 to 5. Other manipulations suggest that verbal and graphical cues also influence scores (Friedman and Friedman, 1994). In addition, people may not be able to assess the difficulty of a task, especially when they are not familiar with the exercise.

Consequently, between sample comparisons may not be reliable. The within sample comparison would seem more appropriate as it relies on intra-individual comparisons. The arbitrariness of the first scores would be compensated by the coherence of the second round scores with respect to the first round scores. People might not know how to interpret the scale or may not be fully aware of the difficulty encountered when completing the task. But when faced with the second scale, they may use it in the same fashion, the scale having been "imprinted".

A condition for the within sample approach to be reliable is that the sequence order does not influence the conclusions drawn. Very often, this criterion cannot be checked since the within sample comparison typically implies the use of one sample only. If paired comparisons show that grouping is more difficult than ranking when presented first but less difficult when positioned second, a within sample comparison might not be appropriate. A way to limit the risk of misleading conclusions is to apply the within sample comparison on two sub-samples, randomizing the order. If conclusions are similar between the two sub-samples, the

confidence on the results may be higher. If conclusions diverge, greater care should be taken when interpreting the results.

Conclusions may diverge when there is a significant learning effect, or when first round scores correspond to a bound of the scale. When there is a sufficiently large learning effect, the second round exercise would be perceived as easier. It is also the case if the task implies a significant fatigue effect, although with the reverse consequence: the second round exercise would be perceived as more difficult. If the learning and fatigue effects are not large enough, or they cancel each other out, the conclusions from the within sample test may remain consistent. If first scores correspond to the lowest bound of the scale, scores cannot decrease whatever the difficulty of the second round exercise. In this survey, although 50% of the participants in each group state 1 at the first round score, the paired comparison lead to similar results in each sub-sample. This suggests that the proportion of participant stating the lowest score needs to be high to affect the overall conclusion.

In this survey, the ranking task is perceived as being significantly more difficult to perform than the grouping task, regardless of the order in which the tasks are undertaken. Conclusions are then independent of the sequence order, thus enhancing their reliability.

7. Conclusion

Rating type questions are sometimes used in valuation questionnaires. One possible application is to assess and compare the difficulty of different valuation tasks. The between sample approach is often privileged for this type of comparison despite its drawbacks. A common alternative is to opt for a within sample approach. Its main advantage is to introduce a clear point of reference according to the coherent arbitrariness principle. One the other hand, the performance of the first task could influence the difficulty of the second one. A way to improve reliability might be to introduce a within sample approach with a split sample to control for succession order effects.

This procedure has been applied in a survey aiming at comparing the cognitive burden of two choice modelling variants, contingent ranking (CR) and contingent grouping (CG). The between sample comparison finds CR and CG being no significantly different in difficulty, while the within sample tests indicate that CR is perceived by respondents as more difficult than CG. The fact that results are independent of the sequence order reinforces this conclusion. In summary, it would seem worthwhile to conduct, where possible, the within sample test estimation, with split samples to control for order effects.

References

Ariely, D., G. Loewenstein and D. Prelec (2003). ""Coherent arbitrariness": Stable demand curves without stable preferences" *Quarterly Journal of Economics* **118**(1), 73-105.

Bateman, I.J. and I.H. Langford (1997). "Budget-constraint, temporal, and question-ordering effects in contingent valuation studies" *Environment and Planning A* **29**(7), 1215-1228.

Brey, R., O. Bergland and P. Riera (2005) "A contingent grouping approach for stated preferences" Working paper number 2005-22, Norwegian University of Life Science.

Brey, R., P. Riera and P.A. Mahieu (2007). "L'intérêt d'utiliser un point de référence pour les questions à échelles dans les études d'évaluation monétaire" *Revue d'Economie Politique* **117**(5), 751-759.

Caparros, A., J.L. Oviedo and P. Campos (2008). "Would you choose your preferred option? Comparing choice and recoded ranking experiments" *American Journal of Agricultural Economics* **90**(3), 843-855.

Flachaire, E. and G. Hollard (2007). "Starting point bias and respondent uncertainty in dichotomous choice contingent valuation surveys" *Resource and Energy Economics* **29**(3), 183-194.

Friedman, L.W. and H.H. Friedman (1994). "A comparison of vertical and horizontal rating scales" *Mid-Atlantic Journal of Business* **30**, 107-111.

Hanley, N., B. Kristrom and J.F. Shogren (2009). "Coherent arbitrariness: On value uncertainty for environmental goods" *Land Economics* **85**(1), 41-50.

Hensher, D.A., J.M. Rose and W.H. Greene (2005). *Applied choice analysis: A primer*, Cambridge: Cambridge University Press.

Louviere, J.J., D.A. Hensher and J.D. Swait (2000). *Stated choice methods: Analysis and applications*, Cambridge: Cambridge University Press.

Mackenzie, J. (1993). "A comparison of contingent preference models" *American Journal of Agricultural Economics* **75**(2), 593-603.

Mitchell, R.C. and R.T. Carson (1989). *Using surveys to value public goods: the contingent valuation method*, Washington: Resource for the Future.

Schwarz, N., B. Knauper, H.J. Hippler, E. Noelleneumann and L. Clark (1991). "Rating-scales - Numeric values may change the meaning of scale labels" *Public Opinion Quarterly* **55**(4), 570-582.

Whynes, D.K., E.J. Frew, Z.N. Philips, J. Covey and R.D. Smith (2007). "On the numerical forms of contingent valuation responses" *Journal of Economic Psychology* **28**(4), 462-476.

Yadav, L., P. Stevens and J. Murphy (2007) "A comparison between the traditional contingent valuation methodology and prediction mechanism" Working Paper, University of Massachusetts.