Bateman, Ian J. et al.

Working Paper

Choice set awareness and ordering effects in discrete choice experiments in discrete choice experiments

CSERGE Working Paper EDM, No. 08-01

Provided in Cooperation with:
The Centre for Social and Economic Research on the Global Environment (CSERGE), University of East Anglia

Suggested Citation: Bateman, Ian J. et al. (2008) : Choice set awareness and ordering effects in discrete choice experiments in discrete choice experiments, CSERGE Working Paper EDM, No. 08-01, University of East Anglia, The Centre for Social and Economic Research on the Global Environment (CSERGE), Norwich

This Version is available at:
https://hdl.handle.net/10419/48801

Choice Set Awareness and Ordering Effects in Discrete Choice Experiments

by

Ian J. Bateman, Richard T. Carson, Brett Day, Diane Dupont, Jordan J. Louviere, Sanae Morimoto, Riccardo Scarpa, Paul Wang

CSERGE Working Paper EDM 08-01

Choice Set Awareness and Ordering Effects in Discrete Choice Experiments

by

Ian J. Bateman*
University of East Anglia, UK

Richard T. Carson
University of California, San Diego, USA

Brett Day
University of East Anglia, UK

Diane Dupont
Brock University, Canada

Jordan J. Louviere
University of Technology, Sydney, Australia

Sanae Morimoto
Kobe University, Japan

Riccardo Scarpa
University of Waikato, New Zealand

Paul Wang
University of Technology, Sydney, Australia

***Corresponding author. Contact details:**

Tel: ++44 (0) 1603 593125
Fax: ++44 (0) 1603 593739
Email: i.bateman@uea.ac.uk

Acknowledgements

The support of the Economic and Social Research Council (ESRC) is gratefully acknowledged. This work was part of the interdisciplinary research programme of the ESRC Research Centre for Social and Economic Research on the Global Environment (CSERGE).

This research was supported by the ChREAM project funded by the ESRC and other research councils under the RELU project (reference RES-227-25-0024). Support from the Economics for the Environment Consultancy (EFTEC) is also gratefully acknowledged.
ISSN 0967-8875

Abstract

The choice experiment elicitation format confronts survey respondents with repeated choice tasks. Particularly within the context of valuing pure public goods, this repetition raises two issues. First, does *advanced awareness* of multiple tasks influence *stated* preferences from the outset, and second, even in the absence of such awareness, does the process of working through a series of choice tasks influence stated preferences leading to choice outcomes that are dependent on the *order* in which a question is answered? The possible motivators of these effects include economic-theoretic reasons such as strategic behavior, as well as behavioral explanations such as response heuristics and learning effects. A case study of a familiar good (drinking water quality) combines a split sample treatment of the presence/absence of advanced awareness with a full factorial design permitting systematic variation of the order in which choices are presented to respondents. A further sample division allows examination of effects arising from variation in the scope of the initial good presented to respondents. Using discrete choice panel data estimators we show that both advanced awareness and order effects exist alongside interactions with the scope of the initial good.

Keywords: discrete choice experiments, strategic behavior, advanced awareness, ordering, WTP, drinking water, mixed logit, heterogeneous preferences.

JEL Codes: Q2, Q25, Q26, C35

1. INTRODUCTION

The discrete choice experiment (DCE) method[1] (Louviere, Hensher and Swait, 2000) has become the most popular approach for valuing a range of multi-attribute public goods. Its origins lie with McFadden's random utility model further developed in the context of using stated preference data in marketing and transportation (*e.g.,* Louviere and Hensher, 1982). From early work focusing upon outdoor recreational activities (Carson, Hanemann and Steinberg, 1990; Adamowicz, Louviere and Williams, 1994), use of DCE has spread rapidly through the environmental economics literature (*e.g.,* Hanley, Wright and Adamowicz, 1998; Bennett and Blamey, 2001; Garrod, Scarpa and Willis*,* 2002). The DCE method, however, has yet to be exposed to the same degree of testing that has been applied to more established preference elicitation methods such as a single binary choice contingent valuation (hereafter SBC) question. While there are some studies comparing willingness-to-pay (WTP) estimates of the same public good from both DCE and SBC surveys of the same populations (*e.g.,* Foster and Mourato, 2003; Scarpa and Willis, 2006; Tuan and Navrud, 2007), formal tests of the DCE approach in a pure public good valuation context are still few (Boyle, Morrison and Taylor*,* 2002; Louviere, 2003). In particular, existing studies do not fully exploit econometrically the main distinguishing feature of choice experiments, namely the panel nature of the discrete responses[2]. This paper is offered as a contribution to this research gap.

One focus of testing the contingent valuation approach to public goods valuation has been to look at the incentive compatibility of various value elicitation formats. The most commonly endorsed (Arrow, *et al.,* 1993) is the SBC format, in which survey respondents face just one question. Carson and Groves (2007) show that this format is incentive compatible provided that the characteristics of the WTP scenario credibly entail: (a) a take-it-or-leave-it proposition on obtaining the good in question at a single specified price, (b) a consequential response in the sense that respondents perceive that the government will take the survey results into account in its decision making, and (c) the government has the ability to compel payment at the stated price if the good is provided.

[1] The DCE elicitation format is referred to under a number of alternative names including choice based conjoint, choice experiments, and choice modeling. In the specific context of valuing public goods, it can be seen as a contingent valuation elicitation format consistent with random utility that is specialized toward valuing differences in the attributes of a good. More generally DCE can be seen as one of a number of stated preference methods, some of which have a clear link to random utility maximization and some which do not.

[2] The DCE format generalizes the standard SBC format by presenting more than one choice question and sometimes by also expanding the number of alternatives available within each choice set from the two considered in the SBC approach. Compared to SBC, the DCE format makes it easier to look at the marginal value of changes in the attributes of a good, while the increase in data generated by multiple responses can improve the statistical power of WTP estimates.

Certain of these conditions for incentive compatibility become compromised within the DCE format which is explicitly designed to confront the subject with multiple choice tasks (thereby clearly contravening condition (a) and arguably undermining condition (c)). The simplest theoretical prediction (Carson and Groves, 2007) in the pure public goods case is that the DCE and the SBC formats will produce different estimates. This is because respondents in the DCE case need to make some type of inference about how responses over multiple questions will be combined to help determine the single level of the public good to be provided so that truthful preference revelation may no longer be an optimal strategy for all respondents.[3] The other main theoretical prediction is that seeing very different prices for the same (or close to the same) good or seeing the same price for quite different goods should lead a respondent to either try to exploit this flexible pricing or to question its credibility. While it is sometimes argued that strategic behavior might not be a problem in DCE due to the complexity of the strategic task demanded of the individual, Carson and Groves (2007) point out that there may be simple strategies that respondents can use that are optimal or close to optimal. This typically revolves around rejecting a preferred alternative in a given choice set when an identical or closely related good was available at a lower price in another choice set.

Strategic behavior may be induced in respondents by knowledge of the possibility of providing multiple responses. We can identify two ways in which individuals become aware of the multiple response nature of a DCE exercise. First, respondents might be made aware of the existence of multiple questions in advance of actually responding. Second, this awareness may arise (or be reinforced) dynamically as respondents progress through the choice questions.[4]

Approaches which permit respondents advanced awareness of multiple choice opportunities (which we will label ADV designs) might take various forms, including informing respondents of the attribute levels, including price, which might be encountered or going further, as is

[3] Since the key pure public good property, that all agents share the same level of the good, is fundamental to the specific nature of the incentives respondents face, it is difficult to infer anything about the incentive properties of the DCE format when used with pure public goods from those using quasi-public or private goods where the incentive structure can be quite different. The DCE format used with public goods provided by voluntary payments may also behave differently due to the change in the incentive structure induced by the payment condition.

[4] This second way is similar to the double bounded dichotomous choice format (Hanemann, Kanninen, and Loomis, 1991). Carson and Groves (2007) note that theoretically the two responses should not be perfectly correlated for all respondents if they take advantage of the strategic incentive and/or utilize the additional informational content of the question. Reasonable additional assumptions yield the stylized fact that the estimate of WTP based on only the first response is higher than that based on both responses.

done in this study, by disclosing the full set of questions they will face. All mail survey DCE studies implicitly allow respondents advanced awareness of questions (although in such circumstances it is difficult to assess the degree to which individual respondents exploit these opportunities). In contrast, in-person and internet DCE surveys have the option of not informing respondents in advance of the multiple task format of the CE exercise, in which case respondents discover this in a stepwise manner as they work through the questionnaire (which we refer to as the STP treatment)[5].

Clearly, a take-it-or-leave-it response between the status quo and a single alternative policy recorded at the start of an STP format DCE exercise is as incentive compatible as the response to a SBC question (indeed they are effectively identical). In a STP DCE survey, this property is eroded as respondents work through the sequence of choices. In contrast, a respondent facing our ADV treatment starts with advanced awareness of the full range of attribute levels and knows that multiple variants of the good are likely to be possible. Hence, even the initial response will not be made under incentive compatible conditions. However, it is possible that the process of working through a series of choice tasks and confronting specific alternatives might reinforce awareness of strategic opportunities even for an ADV respondent. We can, therefore, expect that any deviations from underlying, incentive compatible preference revelation (whether deliberately strategic or otherwise) might be exhibited not only in responses to the first of a series of choice tasks (revealed by contrasting initial responses from the STP and ADV treatments) but also as respondents progress through the choice tasks (observed as an ordering effect). Such behaviors will be revealed within the pattern of responses given by individuals.

The key feature of a pure public good is that only one level can be provided to all people. For any particular provision level a respondent should not say yes to paying an amount that is higher than their maximum WTP for that level because of the risk of having to pay more than it is worth if provided. However, the converse is not true; in order to decrease the likelihood that the least preferred levels of the public good are provided (rather than the more preferred levels), it may be optimal to say no to an alternative that is preferred to the status quo. Carson and Mitchell (2006) looked at reducing the level of risk from low level carcinogens in drinking water using a sequence of open-ended questions to elicit WTP for risk reductions of different magnitudes. They provide suggestive evidence that respondents engaged in strategic behavior in the form of reducing WTP for smaller risk reductions relative to the largest one. We take this as a starting point that suggests that the nature of strategic

[5] These labels follow the Bateman, *et al.,* (2004) investigation of such effects.

behavior may be different for different size goods when multiple competitors to the status quo are possible. Again taking the drinking water example, we look at the effects of presenting in the initial question either a relatively large or small improvement over the status quo. Within the STP (but not ADV) treatment responses to this initial question should be incentive compatible.

The rest of the paper is organized as follows. Section 2 discusses the link between respondent behavior and predicted response patterns in more detail. This is followed by a brief overview of the theory and method used in our analysis. Section 3 notes the most relevant literature. Section 4 of the paper outlines a design specifically developed to address three main research questions. The first question asks whether advanced awareness of the sequence of choice sets to be presented influences initial responses. The second question asks whether the response experience developed through the course of a DCE interview results in systematic changes in stated preferences (ordering effects). The third question asks whether these effects differ by the nature of the good initially presented. Our study is designed to permit such tests.

The sample is first split between the ADV and STP treatments. As noted, a further cross-cutting split ensures that, while some respondents are initially presented with a relatively large improvement over the status quo, others initially face a relatively small good. Respondents in each of these various treatment combinations are then randomly allocated a price level from the vector of cost amounts. While comparisons across the first response provide an important test of awareness (especially given the incentive compatibility of the initial STP response) and any interaction with scope, a straightforward examination of any order effect is provided by taking the initial question seen by a respondent and repeating it as the last question answered. Comparison between these identical questions provides a simple yet effective test of order effects. This is complemented by using an approach that varied the question ordering across a full factorial design permitting us to examine the pattern of ordering effects arising in responses. Section 5 presents results from these analyses. We use a repeated ANOVA approach, as well as various probit models, to look at the effects of choice set awareness on responses to identical first and last choice tasks in the DCE sequence. This analysis clearly establishes the presence of a systematic difference between the results from these two tasks. A second set of results is based on an investigation of the responses to the entire series of choice tasks using panel mixed logit analyses under both finite and continuous mixing of taste intensities. Results suggests that ordering effects can be modeled as a function of the (log of the) order of the sequence at which the response is elicited (interacted with the price variable), implying a WTP which

decreases more rapidly in the first stages of the response sequence than later stages. Section 6 concludes with implications and considerations for future research.

2. BEHAVIOR AND RESPONSE PATTERNS IN DCE

In order to inspect the effects of departure from incentive compatibility within DCE exercises we first need some comparator. Following incentive compatibility arguments, the first response of a DCE respondent facing a pair-wise choice under an STP treatment shares the same incentive compatibility characteristics as a SBC question. We can therefore compare the bid acceptance curves for first responses in the set of DCE choice tasks to assess alternative patterns of behavior under STP and ADV treatments.

The main uncertainty in thinking about the deviation from truthful preference revelation that may arise within the DCE format concerns how the respondent potentially reacts to seeing the same or closely related goods at very different prices. One reaction is to believe that the good can potentially be obtained for a lower price than the one being asked about. We term this the "cost-minimizing" belief and distinguish between the strong and weak forms below. The other reaction is to believe that all of the price amounts seen influence the actual price that would be paid, not just the specific price amount asked about. If the amount being asked about is low relative to the observed range of prices, then the respondent may believe that the true cost will likely be higher and vice-versa. True costs are therefore assumed to be nearer the middle of the price vector under this "cost averaging" belief.[6]

Let us begin by considering the cost-minimizing case. To maximize her surplus, a respondent, holding a WTP for the good which is above the (subjective) expected minimum cost of the proposed policy, will try to secure it at that minimum cost. A likely source for an estimate of the subjective expected minimum cost is the lowest known bid level since a respondent might reasonably surmise that decision makers would only present a project at feasible cost levels. Building upon the classic analysis of Samuelson (1954), if the posted price represented in the DCE exceeds the expected minimum cost, then the strategic incentive is to appear to have a lower WTP and reject any option in the sequence offering it at a higher price than the expected minimum cost. This is the strong cost minimization case.

A weaker but probably more plausible version of the cost minimization hypothesis is that having earlier seen the good (or a similar good) available at a lower price than the one being currently asked increases the perceived likelihood that the good can actually be supplied at

[6] It is possible to give this cost averaging a Bayesian interpretation in which, rather than preferences, it is the expectation of the actual price to be paid for the good which is updated. It should also be noted that this form of cost minimization is equivalent to re-evaluating the quality of the good (downward or upward) in terms of stated price. However, the nature of the good (tap water) involved in this study make this interpretation unlikely.

a lower price than the one being asked about. This provides an incentive to bargain downward by sometimes saying 'no' even when the good at the price asked about is preferred to the status quo.

The difference between the strong and the weak version of the cost minimization hypothesis is that in the strong version the consumer believes there is no risk in not getting the good by saying no to any amount higher than the lowest price the good was offered at.[7] Under the weak version, there is a perceived risk of not getting the good. As the consumer is trying to maximize surplus, the higher the price being asked about (conditional on true WTP being above this price), the more likely the respondent is to say no when truthful preference revelation would have resulted in a yes response. Effectively, the consumer is trading off the risk of not getting the good by saying no at a price that would increase utility against the gain in surplus from obtaining the good at a lower price.[8]

Consider the expected effects of the two choice set awareness treatments (the incentive compatible STP and the incompatible ADV) on the bid-acceptance curve for initial responses in a DCE sequence. If respondents react to advanced awareness of the bid vector by adopting cost-minimizing behavior, then the bid acceptance curve under the ADV treatment will be steeper than under the incentive-compatible STP treatment. That is, it will decline more rapidly at high bid amounts. Assuming that the bid vector, to a large extent, encapsulates the true values of the good (as reflected in the initial response STP curve) then we should also expect the ADV curve to lie below the STP curve at prices above the minimum cost.

Continuing with this scenario but moving along the sequence, a second effect might occur. As STP respondents progress through the sequence of choice tasks they become increasingly aware of the choice set and will begin to revise down their WTP since they realize that there is room for strategizing. Note that this effect will also apply to ADV respondents if they have not fully appreciated the extent of the strategy space from the outset or if they are not able to retain all the information from all the choice sets they saw in short term memory. As such, from the second response onward, we should expect the cost-minimizing behavior to yield an ordering effect consisting of a steady reduction in stated

[7] This behavior is very different from the usual status quo/protest behavior because the respondent indicates a "yes" response to at least one non-status quo alternative.

[8] If: (a) the consumer's surplus (given that the good is supplied) is monotonically decreasing in price, (b) the probability of the good being supplied is monotonically increasing in price, and (c) the consumer says no to any price with a negative surplus, then the response curve under weak cost minimization is equal to the incentive compatible response at the lowest price and lies smoothly underneath it as price increases.

WTP across questions under both choice set awareness treatments. A reasonable working hypothesis is that this effect should be stronger for the STP respondents than the ADV respondents.

A simpler heuristic may lead to the same comparative steepening of the bid acceptance curve as cost-minimizing behavior. As respondents become aware of available combinations of goods and prices on offer, they may increase their rejection rate for combinations offering relative 'bad deals' and accept more frequently those offering relative 'good deals', where the word 'relative' refers to combinations that have already been disclosed in previous choice-tasks. The crucial difference between this 'good deal / bad deal' heuristic and weak cost-minimizing behavior is that, while the latter assumes respondents have well formed preferences at the point of the initial response, the former does not. In other words, the bid acceptance curve can drop as the sequence proceeds without respondents having a clear valuation for the good.

A straightforward test to distinguish between these two strategies can be provided by exposing respondents to a treatment effect related to the nature of the good on offer. Suppose one sub-sample is offered a 'large' good and another is offered a 'small' good. This should have no impact upon the stated preferences of those acting under a "good deal / bad deal heuristic" as they do not rely upon any underlying concept of the value of a good. The test for this sort of behavior is the well-know scope test recommended by the NOAA Panel (Arrow, *et al.*, 1993).[9] In contrast, those with a clear valuation for the good (cost-minimizers) will exhibit different response rates at the same bid levels, thereby revealing a clear difference in stated values between these two goods for the STP treatment in the initial question.[10] Beyond the first question, the "good deal / bad deal" heuristic is what makes engaging in something close to optimal strategic behavior under reasonable assumptions easy to execute.

While the strategic behavior underpinning cost-minimizing responses has a strong theoretical basis, it is not the only pattern of behavior that might be induced by the DCE format's lack of incentive compatibility. Indeed, the main problem with rationalizing a pure cost minimizing strategy is the implausibility that goods of different sizes can have the same

[9] See Carson, Flores and Meade (2001) for a discussion of empirical applications of the scope test.

[10] Note that for clarity, this discussion and those regarding other behaviors tend to describe effects as if they are applying to a single respondent who wholeheartedly adopted the behavior mentioned. In reality, samples of respondents may take on a mixture of behaviors (including truthful preference revelation) with discernable effects being generated by different proportions of respondents adopting different strategies or by respondents adopting some mixture of strategies.

cost and that the exact same good can have different costs. This might lead a respondent into not believing that the cost amount being asked about is indeed the one that would be paid if the good is provided. The most likely form of behavior this would induce is cost averaging. Here, respondents who are given advanced awareness (ADV treatment) of the range of prices to be used in the survey might assume that the actual cost of the project will turn out to lie somewhere in the range bounded by the highest and lowest bids and instinctively identify the cost in the middle part of the range as being more likely to represent the true price that will be charged. The likelihood of having to pay the lowest or highest cost should be the most heavily down-weighted.

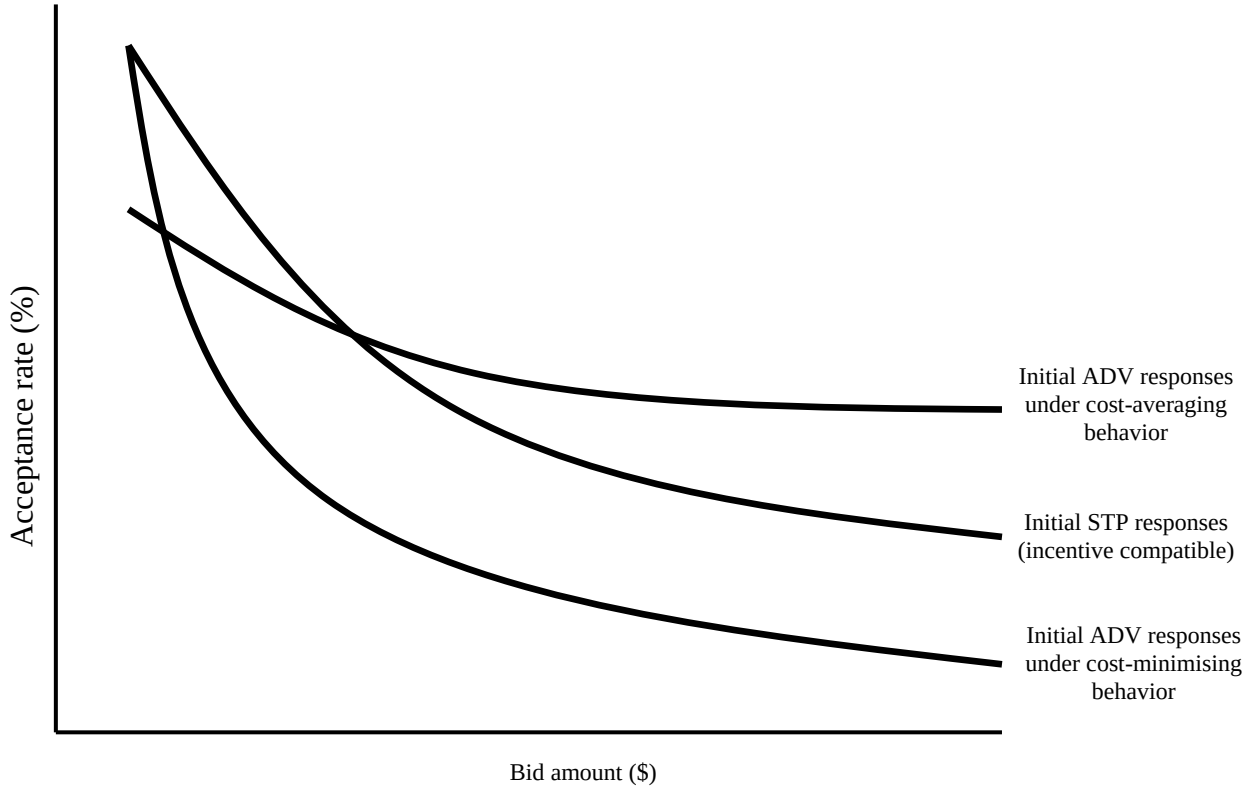
A further refinement on this belief would let the weights on the cost asked about vary with the size of the good. One would expect the effect to be the largest for the highest price matched with a small good.[11] As a consequence, when looking at responses to the first question in a sequence, we might expect to see fewer ADV respondents accepting at the lowest bid amount than we would if respondents believed the stated amount was the actual cost of the project (as per the initial question in the STP sequence). In contrast, if the bid amount is drawn from the high end of the range, then respondents' expectations may be that the actual cost will be lower, leading to a higher acceptance rate than might be expected under cost-minimizing behavior. Under the ADV treatment, such cost-averaging behavior should give rise to a flatter bid acceptance curve for the first question of the DCE sequence than that given by the incentive compatible initial STP response. If the bid vector spans a sufficiently wide range, then the cost-averaging curve should cut the initial response STP curve.[12]

Differences between the cost-minimizing and cost-averaging behaviors for initial responses are illustrated in Figure 1. The former strategy results in a bid acceptance curve for the ADV treatment that is steeper and consistently below the STP curve. In contrast, with cost-averaging behavior, the bid response curve for the first response in the sequence would be steeper for those respondents receiving the STP treatment. Additionally, provided that the bid-vector covers the realistic range for WTP for the good being valued, then the ADV curve is expected to cut the STP curve from below as bid levels increase.

[11] This case represents the maximum surplus extraction by the agency and it is unclear why the agency would reveal in the ADV treatment that it was going to also make an offer of a large good at the same price. The opposite case, a large good for the lowest price, represents minimum surplus extraction but may be a more plausible situation if consumers believe the agency needs strong support to go forward with provision and can actually supply the large good at the lower price. Weights need not therefore be symmetrical.

[12] If the bid vector does not span a sufficiently wide range, it is possible for the ADV curve provided by cost-averages to lie entirely above (small bid amounts) or below (large bid amounts) the STP curve.

Figure 1: Bid Acceptance Curves for Initial Responses from Incentive Compatible (Initial STP) Compared with Cost-Minimising and Cost-Averaging (ADV) Behavior Patterns



While seeing conflicting cost amounts may violate face validity, this conclusion is not inevitable. Most if not all respondents may simply take each amount asked about as the actual cost and respondents may or may not take advantage of the strategic opportunities the DCE format offers. As such, it is useful to test whether cost averaging and/or cost minimization behavior appears to exist.

Even if respondents do not react to the violation of face validity with respect to price with some strategy such as cost averaging or cost minimization, the repetition of tasks inherent in the CE approach may still afford individuals with the opportunity to become more familiar with the nature of choice questions and the contingent market involved (a process termed 'institutional learning' by Braga and Starmer, 2005). More fundamentally, repetition may provide respondents with the opportunity to reconsider and learn about or 'discover' their underlying true preferences ('value learning' as suggested by Plott, 1996). It is possible that both 'institutional' and/or 'value' learning occur within studies allowing repeated contingent

valuation responses (see, for example, Bateman *et al.*, 2008). The usual notion of learning within DCE studies is that it manifests itself by a reduction (increase) in the magnitude of the variance (scale) parameter as the respondent progresses through the sequence of questions or (at least) until fatigue sets in. One implication of the usual learning hypothesis is that it should be faster under the ADV treatment relative to the STP treatment.[13]

[13] Clearly more complex notions of learning might involve substantive systematic changes in preferences parameters over a sequence of questions rather simply an overall increase in the scale parameter.

3. LITERATURE REVIEW, THEORY AND METHODS

Our investigation of advanced awareness of choice tasks and ordering effects within DCE requires two different approaches. The first explores strategic behavior as manifested in differences in bid-acceptance curves. This is analyzed using standard methods for contingent valuation data analysis. The second examines the ordering effect on WTP estimates of a sequence of choice task responses and is based on panel data discrete choice models. We review the relevant theory and applications below.

The statistical analysis of binary responses to take-it-or-leave-it valuation questions is based on random utility theory and is well-known to environmental economists from the large body of studies conducted in referendum contingent valuation (Hanemann and Kanninen, 1999; Carson and Hanemann, 2005). Using choice tasks proposing identical policies, we compare the estimates of bid-acceptance curves for responses observed at the beginning and at the end of the choice sequence, focusing on the effects of choice awareness treatments (ADV and STP). In the statistical analysis we use the standard ANOVA framework, as well as a parametric approach employing probit models. These are used to explain the probabilities of bid acceptance and to retrieve the underlying utility structures. To account for a potential dependence in responses by the same individual (Poe, Welsh and Champ, 1997), we also estimate bivariate probit models to explore the robustness of our results to heteroskedasticity.

The methodological literature on DCE that has focused on issues related to ordering effects begins with the family of multinomial logit models (MNL) anchored on random utility theory. From an economic perspective, this theory postulates that the population is made up of rational individuals who, once faced with set A representing a finite collection of alternatives $A = \{A_1, A_2, \ldots A_J\}$, each systematically chooses the alternative associated with the highest utility. However, the researcher can only observe part of the determinants of the utility that individual n receives from alternative i (the indirect utility or V_{in}), while the other part remains unobserved and represents the idiosyncratic error ε_{in}. Furthermore, for practicality and because any function can be locally approximated by a linearized expansion, indirect utility is often assumed to be linear in K fixed marginal effects (or taste-intensities):

$$U_{in}=V_{in}+\varepsilon_{in} = \sum_{k=1}^{K} \beta_{ik} x_{ik} + \varepsilon_{in} . \qquad (1)$$

If the error term is independently and identically distributed according to a Gumbel distribution (with fixed scale λ inversely proportional to the common variance of the error (or

$\lambda = \pi\,(6\sigma^2)^{1/2}$, where σ is the standard deviation of the Gumbel error), then the probability of selecting an alternative *i* from the common pool of *J* is given by the convenient logit specification:

$$P_{in} = \frac{e^{\lambda V_{in}}}{\sum_{i=1}^{J} e^{\lambda V_{jn}}}. \qquad (2)$$

If the errors are normal, this leads to a similarly defined probit model. The binomial case used in referendum contingent valuation is the specific case in which, for one alternative, one exponential is set to 1. Note that in this specification, scale does not vary across respondents and it cannot be identified separately from the vector of taste of fixed intensities, β, which is taken as equal to one. However, ratios of scale parameters can be identified by fitting different scale parameters to separate datasets (Swait and Louviere, 1993).

As the indices *i* and *n* indicate, the variables used to parameterize the scale may now depend on the alternative and on the individual. Models using this specification are often called the 'heteroskedastic logit' because this parameterization implies that the error variance also depends on the selected parameter vector α via equation (2). The variables typically used in *z* are socio-economic characteristics of the respondent that are thought to be related to the respondent's ability to cognitively perform the task (Scarpa, *et al.,* 2003) or descriptors of choice complexity and other features of the experimental design (*e.g.,* Dellaert, Brazell and Louviere, 1999; DeShazo and Fermo 2002; Caussade, *et al.,* 2005).

Modern analysis of panels of multinomial responses is based on generalizations of the multinomial logit model above, with the most commonly used being the panel data mixed logit. These are basically multinomial logit models integrated over mixing distributions of varying, rather than fixed, parameters. They often do not have a closed-form integral which necessitates the use of simulated maximum likelihood techniques (Train, 2003). Distributions can be continuous or finite (with degenerate mass at given probability points). The latter are often called latent class logit models

Previous efforts to examine patterns of response behavior in choice experiments have looked at how a question framework affects responses[14] and have, with very few exceptions,

[14] These include: optimal experimental design relating to number and ordering (Farrar and Ryan, 1999) of attributes in choice tasks, number of attribute levels and correlation between levels within a choice task (DeShazo and Fermo, 2002; Phillips, Johnson and Maddala, 2002), number of choice tasks (Johnson and Orme,

assumed that respondents' stated preferences are unchanging throughout the choice task sequence. One exception is provided by Carlsson and Martinsson (2001) who use a design with sixteen non-identical choice tasks: eight of which are put into set A and the remainder into set B. They examine the stability of preferences by comparing estimates for observed choices in subset B, when presented as the first eight choice sets encountered by respondents as opposed to when presented as the last, and to preference estimates obtained by pooling all observations. Neither the null hypothesis of equal taste-intensity parameter, β, nor the hypothesis of equal scale, λ, is rejected, although a relatively small sample of 315 choices from a total of 35 students is employed. Phillips, Johnson and Maddala (2002) employ a similar format (pair-wise choice sets without a status quo alternative) with identical choice tasks placed early and late in the survey. The authors note that, in the majority of cases (77%), the same choice is made in both tasks. Under the traditional econometric view of a random utility model, one would expect to see the same choice made 100% of the time. However, under the (Thurstone) psychology view of a random utility model there is a true random component, so one would expect to see some deviations. As such, testing whether the preference parameters have changed may be a better test from the perspective of economic relevancy.

Observing systematically different responses to identical choice tasks when placed at different points throughout a sequence may suggest the presence of changing stated preferences. However, this snapshot-in-time approach does not reveal the nature of dynamic change or ordering effects taking place as respondents make progress through multiple choice tasks. For example, as previously noted, one might speculate that the DCE format offers the opportunity for learning along the lines discussed by Plott (1996). Swait and Adamowicz (2001) believe that the first choice tasks in a sequence of DCE questions may provide a way to learn about the task format, including the amount of effort needed to provide answers and consideration of response strategies. However, they express concern that later choice tasks, rather than affording further learning opportunities, may induce fatigue or boredom (although this depends on the subject matter and complexity of the tasks). The test typically used to distinguish between such effects is based on the argument that learning implies a smaller noise to signal ratio from observed choices. Conversely, fatigue manifests itself as a larger noise to signal ratio (by an increasing value for the variance of the Gumbel error term (ε) in later choices, or equivalently, by decreasing its scale).

1996), and range of values for attributes in choice tasks (Dellaert, Brazell and Louviere, 1999; Ohler, *et al*., 2000).

Phillips, Johnson, and Maddala (2002) find evidence of fatigue in a split sample comparison of choices made at the beginning (the first six questions) and at the end (the last six questions) of a choice task sequence. On the other hand, Adamowicz, *et al.* (1998) use eight choice tasks and find no evidence of fatigue as measured by an increase in variance of the Gumbel distributed error-term. Holmes and Boyle (2005) use a sequence of four choice sets based on binary responses and examine responses from each choice task in the sequence by fitting four separate binary probability models. They find that the relative scale parameter for the fourth and final choice task is significantly greater than the relative scale parameter for the initial question, which is taken as evidence of learning. [15] Swait and Adamowicz (2001) use latent class models (LCM) to look at complexity and task order as a proxy for learning and fatigue effects. They argue that, as task order increases, respondents move from rich, attribute based strategies to simpler ones that focus upon key attributes, such as price and/or brand information. Finally, Caussade, *et al.* (2005) in their 'design of designs' study find that scale gradually increases along the sequence to peak at around nine to ten choice tasks and then decreases.[16]

Something not addressed in these papers is the best way to model order effects. We discuss our approach using both a panel mixed logit approach and a latent class modeling approach in Section 5.

[15] Holmes and Boyle (2005) conduct their DCE using a mail experiment where it is possible for the respondent to see the complete set of choice questions. In some ways this is similar to our ADV treatment. They show that both the preceding and following questions influence the choice made from a particular choice set. Scale also does not monotonically increase over the four questions which is inconsistent with a simple progressive learning story.

[16] Although we follow the standard approach here of interpreting decreases (increases) in variance (scale) as implying more precise, better quality estimates, a strong cautionary note is needed. In the standard OLS model reductions in the variance parameter imply better quality predicted values. However, in choice models the variance parameter is tightly linked to the estimate of market shares, the larger the variance parameter the more equal the estimated market shares are and the smaller the variance parameter the more unequal the estimated market shares are. As such, it not clear whether a large or small scale parameter is better for making predictions about future market shares (Louviere, *et al.*, 2002).

4. DESIGN AND SURVEY DESCRIPTION

4.1 Design Protocol

To further the methodological objective of our study, we employed a good which is familiar and easy to explain to respondents; the quality of domestic drinking (tap) water. Tap water in the United Kingdom is considered amongst the best in the world in terms of consistency of supply and health risk levels due to a combination of naturally occurring factors (*e.g.,* rainfall, temperature) as well as high investment and monitoring levels (Hunter, 2002). Surveys of customer opinions find that the main areas of concern for most consumers are largely confined to issues of water discoloration, odor and taste problems (MORI, 2002). Additionally, climate change and in-migration to the study area are likely to impinge on these aspects of water supplies in the future (Kabat, *et al.,* 2002; Holman, *et al.,* 2002)[17]. Thus, we developed scenarios that varied water discoloration and odor/taste attributes of water quality. We used focus groups to refine the scenarios in the DCE survey. Specifically, we found that water supply states can be described by three attributes: (1) the number of days annually where a household's tap water smelled and tasted of chlorine (ST)[18], (2) the number of days annually where the tap water was a rusty color (C) and (3) the addition to the household's annual water bill induced by implementing technical procedures to address these problems.

The choice tasks in the DCE survey used the simplest response format, namely a choice between a constant 'status quo' (SQ) and a single 'alternative' state, varied across choice questions (thus ensuring that the initial response in the STP treatment is identical to a contingent valuation SBC question). The SQ was defined as the likely level of tap water problems to be experienced over the coming year. Focus group tests indicated that credible maximums were 10 days per year for the ST attribute and 5 days per year for the C attribute in the absence of any intervention (a zero increase in the water bill) to address these problems. We assigned four levels to each attribute: ST (0, 3, 6 and 10 days); C (0, 1, 3, and 5 days), and water bill (£10, 20, 30 and 50). All possible combinations of these three attributes is the full factorial, or $4^3 = 64$ combinations (Louviere, Hensher and Swait, 2000).

We constructed two designs from the three attributes each of which has four levels: (1) the full 4^3 factorial, and (2) a 2^3 (=8) involving only the two extreme levels of each attribute. Although full factorial designs are relatively rare in environmental applications (exceptions

[17] We are grateful to Irene Lorenzoni (UEA, UK) for conversations regarding public perceptions of climate change.

[18] Incidents of smell and taste problems were too collinear to be separated, so were treated as a single attribute.

are Bullock, Elston and Chalmers (1998) and Garrod, Scarpa and Willis (2002)), they allow the analyst to estimate all main and interaction effects independently of one another. In addition, we used a Latin square design (Street and Street, 1987) to control for order effects in association with the smaller factorial. That is, the 8 scenarios produced by the 2^3 were assigned to each order position in a balanced manner using a Latin square. Since 64 combinations were felt to be too many for any individual respondent to cope with, these combinations were divided into four equal blocks, yielding a reasonable number of 16 choice tasks per respondent. Each of those blocks received the same complete factorial design of the extreme levels plus 7 additional scenarios drawn from the 64 possible scenarios. Half of the respondents received the block of only extreme levels first and the other half second. Adding in the identical first and last choice (concerning either the 'large' or 'small' good with the price randomly drawn from the vector above) discussed previously meant that, in total, each respondent faced 17 choice tasks.

Our design allows us to examine the cumulative impact of order related effects over the sequence of choices and, by splitting the sample, to look at differences in such effects between the ADV and STP treatments. Here strategic behavior related to a single attribute (like price) may progressively increase the absolute value of that attributes' coefficient as respondents work through the question sequence. However, this is confounded with the estimate of scale differences.[19] We can address this by looking at changes in scale as a function of sequence order in a way that improves upon the existing literature but we have to caution that teasing out such effects is statistically more fraught with difficulties than generally assumed. The major difficulty here is that sequence order effects are generally confounded with the particular mix of choice sets respondents saw at that order in the sequence unless the assumed data generating process is the true one.[20] Ideally, one would like to interpret sequence order effects as having been averaged over the set of possible designs rather than having possible order effects confounded with a particular design. We argue that our design does a better job of approximating this than existing studies but there

[19] To see this, note that changing the absolute value of any single coefficient across choice sets increases the signal to noise ratio in the estimated model in a way that is indistinguishable from reducing the variance estimate.

[20] To see this, consider the case where a simple linear utility function is fitted to the data when true utility function has some curvature. If the mix of choice sets that respondents see contains a limited range of attribute levels then it would be impossible to distinguish between these two utility functions; whereas, if the mix of choice sets seen contained a broader range of attribute levels, then a different set of parameter values would have be obtained. This can be seen as a specific example of the issue pointed out by Swait and Louviere (1993) when they note that specification problems can get reflected in the variance estimate. Even if the correct data generating process is assumed for all respondents, the precision of the estimates depends upon the mix of choice sets observed.

are some divergences of which two are worthy of consideration.[21] The first divergence from an ideal design is inevitable given our purpose of comparing the same first and last choice and using only two alternatives (the 'large' and 'small' good) to assess the scope interaction issue: the covariance matrix is necessarily limited relative to a balanced covariance matrix using all attribute levels. Correcting the second issue, which stems from how respondents were randomly assigned to choice sets in orders 2 through 16 is possible but requires a more complex experimental design than used here. While these issues have been ignored in previous studies looking at order effects, they may influence the interpretation of our later results in Section 5.3 and should be taken into account in future studies.

The choice task awareness issue was addressed by randomly allocating respondents into either an STP or ADV treatment. Common to both treatments was a section of the questionnaire which described the nature of the two non-price attributes and the underlying sources which caused them without mentioning specific levels. Respondents receiving a STP survey were presented with choice tasks in the typical sequential manner, that is, the choice response for one task is elicited prior to the presentation of a subsequent choice task, thereby involving no prior awareness of forthcoming choice tasks. In contrast, within the ADV treatment and prior to the initial choice task, we provided respondents with details of all attributes and levels used in the study. This was achieved by first showing respondents the text show card reproduced as Figure 2. These respondents were then shown, in the order of presentation, each of the questions which they would subsequently face during the choice exercise. Only then were ADV respondents asked to answer the initial choice question.

[21] Existing studies tend to use a less than full factorial design with no or minimal randomization across sequence order. We use a full factorial, a Latin Square design to control for the ordering of choice sets taken from the 2^3 part of the full factorial, random assignment to blocks from the larger 4^3 design and random assignment of the order of these two components. Still the implementation was inadequate to ensure equivalent covariance matrices across the sequence orders. In particular, orders 2, 3, 5, 10, 11 and 13 are missing one of the interior water quality attributes, suggesting that finer control over blocks taken from the 4^3 is needed to achieve a more complete balance of attribute levels. The most severe problem is at sequence order 9, which is the join point between the 8 choices from the 4^3 design and 7 from the 2^3 design, which is missing the middle levels of all attributes. This issue could have been avoided by having 8 rather than 7 choice sets drawn from the 4^3 design. The original design assumed the first and last choice would be taken from the 2^3 extreme alternatives. This was later changed to gain more power by reducing the number of initial choice alternatives to two to gain power and by changing those choices to non-extreme alternatives to enhance the sensitivity of the scope test. Unfortunately, an additional choice set from the 4^3 design was not added to the overall design when this was done to balance the 9^{th} order. Comparisons between the first and last choice set do not involve these issues as they share the same covariance matrix and the cumulative mix of choice sets seen between orders 1 and 17 is statistically equivalent for all of the treatments.

Figure 2: Inducing Prior Awareness Choice Set: Show Card Presented to ADV subjects

In these questions you will be presented with some schemes which reduce the number of days of chlorine smell and taste from the maximum of 10 days to either 6, 3 or 0 days.

You will also see schemes which reduce the number of days of rusty coloured water from the maximum of 5 days to either 3, 1 or 0 days.

However, you will also see different additions to your water bill for the costs of these various schemes ranging from £10, to £20, £30 and up to a maximum addition of £50 per year.

You will also see combinations of all of these.

Note: The original show card used font 26 bold, two-colored text to aid easy reading of the text and ready understanding of the attributes and levels entailed. Recall that in addition to this information, each ADV respondent was also shown all of the questions they would face in advance of any response being elicited.

For the initial choice question, respondents were randomly allocated to either the relatively 'small' improvement over the status quo or the relatively 'large' improvement (however, these labels were not disclosed to respondents). Table 1 describes the attribute levels defining the two goods and contrasts this with the status quo levels of these attributes (recall however, that each choice question only presents one alternative to the SQ). In addition to these distinctions, we further randomly assigned respondents to one of the four price levels shown.

Table 1: Definitions of the 'Small' and 'Large' Improvement Goods

	Status Quo (SQ)	Only one presented per sub sample	
		Small Improvement	Large Improvement
Number of days each year on which your tap water smelled and tasted of chlorine	10	6	3
Number of days each year on which your tap water was a rusty color	5	5	3
Addition to your annual water bill *(only one presented per subgroup)*	£0	£10, £20, £30, £50	£10, £20, £30, £50

By combining the STP versus ADV and 'small' versus 'large' distinctions, we have four sub-samples to which respondents were randomly assigned for their initial choice question (this being repeated as the last question each respondent faces). A further split of each sub-sample by one of the four price levels allowed us to examine the bid-acceptance curves for the first choice task responses in light of the cost-minimizing versus cost-averaging behavior patterns discussed previously.

4.2 Survey Implementation and Sample Characteristics

A team of experienced interviewers that were specially trained in how to administer the survey conducted face-to-face interviews at respondents' homes. Respondents' addresses were selected on a randomized basis from an area in and around Norwich (UK) and were designed to ensure variation across household socio-economic and demographic characteristics. The different versions of the survey were randomly assigned to respondents and a sample of 864 completed questionnaires collected.

Aside from the choice tasks, survey respondents were asked to provide information concerning a variety of demographic and socioeconomic characteristics, a selection of which is presented in Table 2. Each sub-sample exhibits substantial variation across these characteristics, however, the socio-demographics statistics are statistically not dissimilar across the four sub-samples as shown by the reported *p*-values.[22] Further questions ascertained that, in line with national trends (MORI, 2002) and our focus group findings, issues of taste and odor constituted the most frequent water supply problems which respondents experienced. Partly as a result of this (as well as convenience) over 90% of the sample purchased bottled water at least occasionally. Respondents reported average water bills of roughly £280 per annum.

[22] Although our survey did not explicitly set out to capture a representative sample of the region, comparison with official statistics suggests that the sample is reasonably representative. For example, regional mean income levels are roughly £25,000 (National Statistics, 2001), a level similar to that of our sample.

Table 2: Sub-Sample Characteristics[1]

	ADV Large good	STP Large good	ADV Small good	STP Small good	*p*-value
	n=242[2]	n=250	n=185	n=187	
Education					
% Secondary School or less	0.32	0.37	0.28	0.26	0.096
	(0.47)	*(0.48)*	*(0.45)*	*(0.44)*	
% Upper Secondary	0.18	0.17	0.16	0.23	0.299
	(0.39)	*(0.37)*	*(0.36)*	*(0.42)*	
% University Degree/Equivalent	0.20	0.16	0.20	0.24	0.231
	(0.40)	*(0.37)*	*(0.40)*	*(0.43)*	
% Professional Qualification	0.30	0.31	0.35	0.27	0.393
	(0.49)	*(0.46)*	*(0.48)*	*(0.45)*	
Income (£ p.a.)	22,731	24,343	24,936	22,901	0.400
	(14330)	*(14554)*	*(14895)*	*(14519)*	
Age					
% 25 and Under	0.12	0.12	0.17	0.19	0.091
	(0.32)	*(0.32)*	*(0.37)*	*(0.39)*	
% 26-45	0.27	0.29	0.34	0.26	0.453
	(0.45)	*(0.45)*	*(0.48)*	*(0.45)*	
% 46-65	0.32	0.30	0.31	0.35	0.735
	(0.47)	*(0.46)*	*(0.46)*	*(0.47)*	
% Male	0.42	0.42	0.45	0.51	0.270
	(0.50)	*(0.50)*	*(0.50)*	*(0.50)*	

Notes:
1. Aside from p-values, cell contents are variable means (with standard deviation in parentheses).
2. More respondents received the Large Improvement treatment than the Small Improvement one. This decision was made on the basis of initial testing that suggested more precision was needed in the Large Improvement scenarios relative to the Small Improvement scenarios.

5. RESULTS

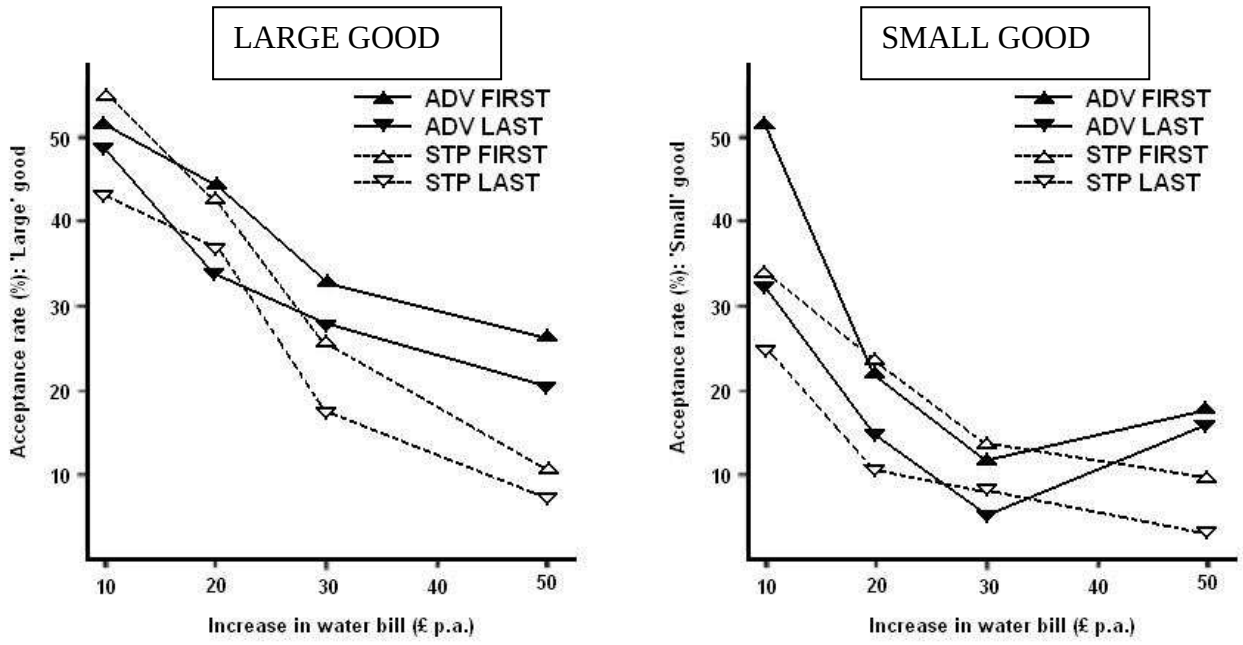
The experimental economics literature suggests a simple approach for examining effects arising as respondents progress through the sequence of choice tasks involved in the DCE exercise. By the simple expedient of posing the same choice task to respondents twice, once at the beginning and once at the end of the sequence, one can use a repeated-measures ANOVA of these first and last responses to reveal any ordering effects. Arguably, this test might under-estimate the extent of such ordering effects if respondents realize that this is the same question and then apply an internal consistency rule which results in them giving the same response. However, if the number of intervening questions is substantial, as in our case, then such deliberate strategies become less likely.

While ANOVA-style testing is intuitively appealing because of the weak assumptions it imposes, most DCE studies adopt stronger parametric assumptions in testing for response characteristics. This has two advantages. First, by bringing more structure to the specification, one can reveal not only the direction but also the rate at which any ordering effects occur. Second, it allows an investigation of ordering effects upon variance. Consequently, we supplement our ANOVA testing with detailed parametric analysis of the entire sequence of DCE responses using probit analysis. We also provide a parametric analysis of the different pattern of responses to the same question posed at the beginning and at the end of the DCE sequence using binary probit analysis under the assumption of separate and joint (bivariate) valuations.

5.1 Visual and Repeated ANOVA Analysis of Responses to First and Last Choice Tasks

Our analysis starts by looking at responses to the (identical) first and last choice tasks in the most intuitive way: visually. The left hand panel of Figure 3 plots bid acceptance response curves for respondents allocated to the ‘large’ improvement good in their first (and hence, also last) choice task. The right hand panel plots corresponding results for those allocated to the ‘small’ good. Within each panel we show four separate response curves (ADV first, ADV last, STP first, and STP last).

Figure 3: Bid Acceptance Rates for Large and Small Improvements



We begin our visual inspection of results (Figure 3) by assessing the choice set awareness effect, starting with the 'large' improvement good. The first thing to note is that the difference between the bid acceptance curves for the large and small goods for the first STP response suggests that respondents are not simply adopting a 'good deal/bad deal' heuristic. A second result is that the strong version of the cost minimization hypothesis is clearly rejected for both the large good and the small good in the ADV FIRST treatment since a non-trivial number of respondents have accepted cost amounts over £10. Next we compare the first responses of the ADV and STP samples ('ADV FIRST' and 'STP FIRST') and observe a clear effect. At the highest bid amount the acceptance rate of ADV respondents is considerably higher than that of their STP counterparts for both goods[23]. This pattern of responses between the ADV and STP treatments clearly does not accord with even the weak version of 'cost-minimizing' behavior. The shallower slope of the ADV curve fits well with the predictions of cost-averaging behavior and it bisects the STP curve as would be expected under that hypothesis (with a wide enough range of bid levels, as evidenced here). The upturn in the ADV curve at the highest price level for the small good is strongly

[23] Note that this need not always be a consequence of cost-averaging. If the price vector was everywhere set below the true average value then cost-averaging would lower mean WTP.

indicative of cost-averaging behavior that puts little weight on the highest amount when asked.[24]

Figure 3 also provides a first insight into ordering effects through a comparison of bid acceptance rates obtained from the first and last questions (ADV FIRST versus ADV LAST and STP FIRST versus STP LAST). Recall that, in each case, the first and last choice tasks in the sequence are identical. It is obvious here that, regardless of treatment type, there is a very clear reduction in bid acceptance rates between the first and the last choice sets of the CE exercise. Contrary to the choice set awareness effect, which fits the cost-averaging model, this reduction in bid acceptance as choice set awareness progressively arises in the STP treatment or is reinforced in the ADV treatment appears to resonate more with weak cost-minimizing behavior.

While these effects appear substantial, we need to establish their statistical significance. Since even in this first versus last comparison there are two observations for each respondent, the data can be analyzed using repeated-measures analysis of variance (ANOVA) procedures (Crowder and Hand, 1990). A repeated-measure design typically contains more than one source of error (within subjects as well as between subjects) and there is a correlation in the within-subjects (repeated) observations. In our design, we have one within-subjects factor representing the placement of a choice task (first versus last) and two between-subjects factors representing the Small/Large Improvement distinction and the STP/ADV treatments.

Table 3 summarizes results from this analysis. The choice set awareness effect, as represented by the STP/ADV factor is statistically significant ($F=4.69$, $p<0.05$), although it is smaller in magnitude than either the size of improvement or the price effect. Other things being equal, when compared to STP respondents, ADV respondents were significantly more likely to choose the alternative scheme over the SQ. Choice set awareness, therefore, has a significant impact upon stated preferences. However, the interaction term between STP/ADV and Small/Large Improvement is not at all significant ($F=0.02$, $p>0.05$), suggesting that the two factors affect the acceptance rates separately.

[24] Indeed, this behavior violates the expected weak monotonicity of the bid acceptance curve, although it should be noted that such violations may occur simply due to sampling variability, something which cannot be rejected statistically in this case.

Table 3: Repeated-Measures ANOVA Results

Tests of Between-Subjects Effects

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Squares*	*F-test value*	*p-value*
STP/ADV	1.38	1	1.38	4.69	0.031
STP by Size of Improvement	0.01	1	0.01	0.02	0.882
Small/Large Improvement	8.86	1	8.86	30.06	0.000
Overall price effect	21.57	3	7.19	24.38	0.000
Price1 covariate	17.36	1	17.36	58.88	0.000
Price2 covariate	3.24	1	3.24	10.99	0.001
Price3 covariate	0.04	1	0.04	0.13	0.720
Residual	252.69	857	0.29		

Tests of Within-Subjects Effects

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Squares*	*F-test value*	*p-value*
Order	2.61	1	2.61	41.69	0.000
STP by Order	0.01	1	0.01	0.12	0.725
Small/Large Improvement by Order	0.02	1	0.02	0.38	0.537
STP by Small/Large Improvement by Order	0.01	1	0.01	0.15	0.697
Residual	53.86	860	0.06		

The other results given in the upper part of Table 3 strongly conform to prior expectations. We observe a highly significant F-test statistic for the Small/Large Improvement factor ($F=30.06$, $p<0.01$) indicating that, as expected, the Large Improvement in tap-water quality results in a significantly higher bid-acceptance rate than the Small Improvement. Consequently, we reject the hypothesis that respondents do not have underlying preferences (which is implicit in the 'good deal / bad deal' heuristic). As expected, we find a highly significant price effect (higher water bill) on choices. This is detailed through the three price variables, Price1, Price2, and Price3, which represent the linear, quadratic, and cubic aspects of the price factor, respectively. Both Price1 and Price2 are highly significant ($p<0.01$) while Price3 is not.

The lower part of Table 3 shows the test results of within-subject effects. The Order factor represents the main-effect of moving from the initial to final response. This effect is highly

significant (F=41.69, p<0.01). This suggests that respondents are significantly less likely to choose an identical alternative when it is presented to them last in a sequence, as opposed to first (all interaction effects proved insignificant). Implicitly, the marginal utility associated with an otherwise identical alternative is lower when it appears at the end of a sequence of choice tasks. Moreover, these results provide support for the hypothesis that such an ordering effect is common to all sub-samples (both STP and ADV and the Small/Large Improvements). This in itself is interesting. While we might expect that those in the STP treatment would react in some way to the increasing awareness of the choice set as they work through the CE questions, it is clear that the ADV respondents are responding with a similar directional effect.

5.2 Probit Analysis of Responses to Identical First and Last Choice Tasks

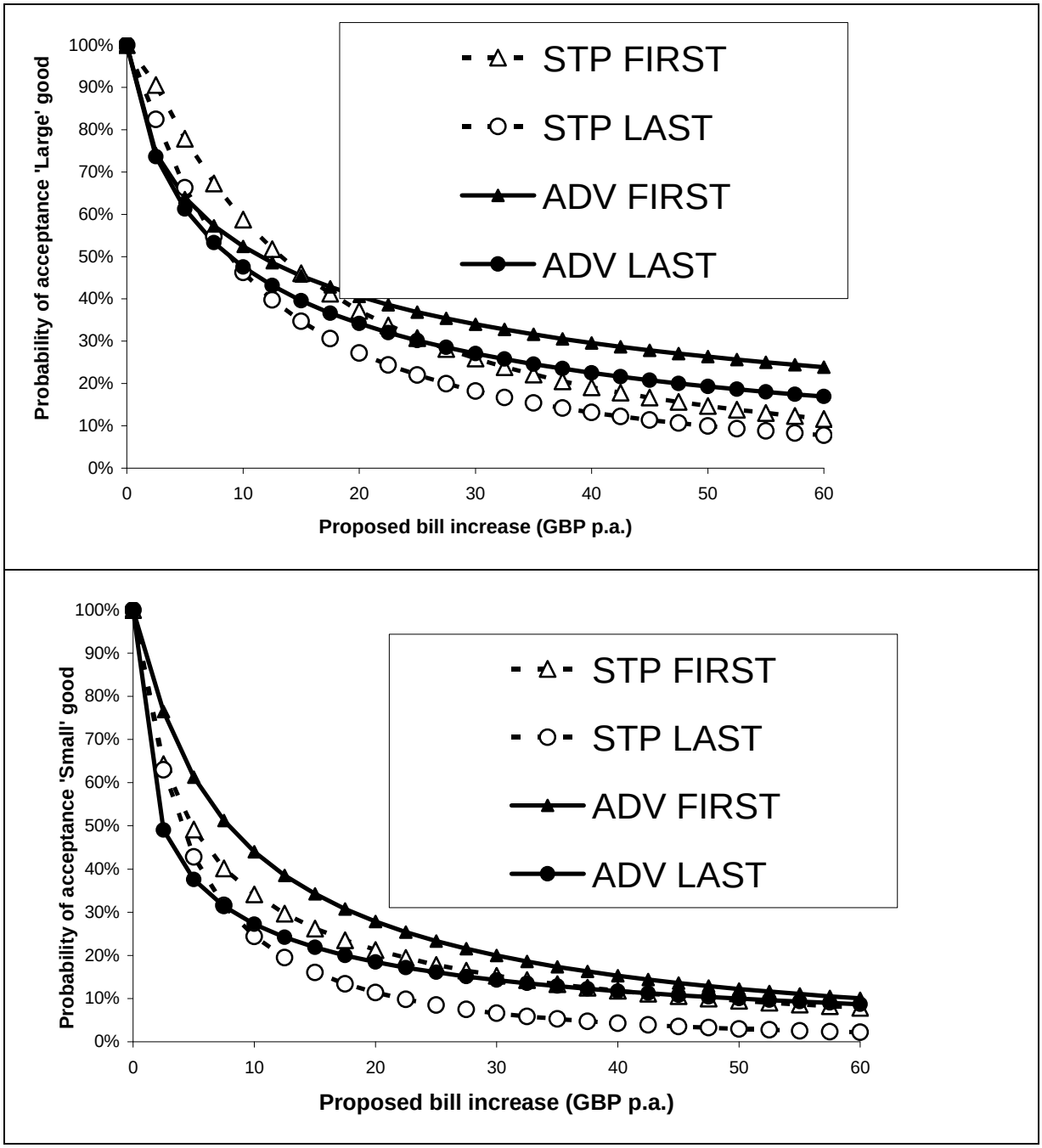
We also analyze the responses to the first and last choice tasks by fitting probit models to the relevant subsets of the data.[25] Table 4 presents estimates for the probability of a yes response for the eight separate sub-samples. Then, within each of these we identify the groups of first (last) responses. Finally, we further split the data according to the choice awareness treatment (ADV/STP). The models include only constant (α) and slope (β) parameters, where the explanatory variable is the log of the bid amount. This permits a plotting of the probability of acceptance curves (Figure 4). From the plots, there is clear evidence for the large good that the STP and ADV curves cross as predicted by the cost averaging hypothesis. The evidence for the small good is suggestive but not as clear cut as the large good case.

Table 4: Binary Probit Results for Comparison of the First and Last Responses

	LARGE GOOD					
	CONSTANT	|*z*-value|	ln(BID)	|*z*-value|	LL	obs
STP FIRST	2.04	4.48	-0.79	5.36	-146.94	250
STP LAST	1.61	3.43	-0.74	4.76	-133.25	250
ADV FIRST	1.05	2.32	-0.43	2.96	-157.65	242
ADV LAST	1.09	2.36	-0.50	3.35	-147.85	242
	SMALL GOOD					
STP FIRST	0.88	1.50	-0.56	2.96	-88.48	187
STP LAST	1.01	1.47	-0.74	3.21	-60.08	187
ADV FIRST	1.30	2.45	-0.63	3.70	-98.81	185
ADV LAST	0.36	0.64	-0.42	2.34	-82.42	185

[25] A probit model is used here because of the ease of extension to the bivariate case which we consider next. Logit models provide very similar estimates.

Figure 4: Bid Acceptance Curves Implied from Separate Probit Models



The results in Table 4 do not account for the obvious dependence between first and last responses from the same individual. Following Poe, Welsh and Champ (1997) we also estimate a number of bivariate probit models that simultaneously explain the probability of acceptance of first and last responses. We estimate three specifications that include dummies for all the effects of relevance. These estimates are reported in Table 5. Starting with a basic bivariate probit that does not account for heteroskedastic preferences (Slope-effects only), we see that coefficients on the log bid variable are negative and significantly different from zero. The interaction effect between the bid and either the size of the good

(LARGE dummy) or the choice awareness (ADV dummy) are both significantly different from zero and positive, as expected. The correlation parameters accounting for the dependence between initial and final response by the same person are very large and highly significant, effectively ruling out independence.

We also wish to allow for the possibility that either the size of the good or the choice set awareness treatments, or both, may affect the error variance instead of the utility function. To do this, we fit two bivariate heteroskedastic models. In the first we maintain a differential slope effect for the large good (which proves highly significant, as one would expect) and account for a multiplicative scale in the error for the ADV treatment with the specification $\sigma_i^2=\exp(\gamma ADV)$, for responses *i*=initial, last. A likelihood ratio test suggests that the addition of ADV in the variance term is insignificant (p=.31) even though the variable itself is significant. In the second, both the size of the good and the ADV are allowed to affect the variance with $\sigma_i^2=\exp(\gamma_1 ADV+\gamma_2 LARGE)$. Again the likelihood ratio test for the joint significance of these two variables is insignificant (p=.29). The BIC values suggest the model that adds the ADV and LARGE indicator variables in the heteroskedasticity function is marginally better.

Point estimates of mean and median WTP implied by separate probit models are presented in Table 6. The test of primary interest is between the median and mean WTP for large and small goods for the STP initial choice. Here the statistical equivalence of the median (p=.004) and the mean (p=.015) are clearly rejected. What is interesting to note is that the statistical significance of the difference between these two summary statistics for the large and the small goods falls for the comparison of the last STP choice, although one would still reject scope insensitivity. The opposite relationship is found with the first ADV response. Here one would accept the null hypothesis of no scope sensitivity whereas with the last ADV response, one would reject it at the .10 level but not the .05 level.[26] One possibility for this reduction in the apparent sensitivity to the scope of the good being valued with the initial ADV treatment is that, in addition to cost averaging, there may also be some initial averaging of the other attributes going on, reducing the apparent difference between the small and the large good. Asking questions about different sized goods may help to sharpen the differences between the attribute bundles as the respondent progresses through the sequence.

[26] The entire STP or ADV DCE always strongly rejects scope insensitivity. Due to the larger number of observed choices, the DCE has tighter confidence intervals relative to looking at either the first or last choice alone.

Table 5: Bivariate Probit Results for Comparison of the First and Last Responses

Model	Slope-effects only				Heteroskedastic + Slope-effects				Heteroskedastic + Constant-effects			
Response	First		Last		First		Last		First		Last	
	Estimate	|*z*-value|	Estimate	|*z*-value|	Estimate	|*z*-value|	Estimate	|*z*-value|	Estimate	|*z*-value|	Estimate	|*z*-value|
Constant	1.39	5.62	1.12	4.24	1.62	5.32	1.34	4.10	1.81	5.84	1.57	4.71
LOG_BID	-0.72	8.63	-0.76	8.21	-0.81	7.59	-0.85	7.18	-0.71	6.94	-0.72	6.40
LOG_BID × ADV	0.06	1.94	0.09	2.71								
LOG_BID× LARGE	0.13	4.38	0.18	5.28	0.15	4.28	0.21	5.01				
LARGE									-0.49	4.36	-0.65	5.17
Heteroskedastic Effect of ADV					0.31	2.08	0.35	2.92	0.31	2.12	0.34	2.83
RHO	0.89	42.47			0.89	42.40			0.89	42.37		
LL	-751.56				-751.04				-750.31			
BIC	1,551.26				1,550.31				1,548.77			

Looking at Table 6, one gets the impression that in the STP treatment respondents reduced WTP amounts between the first and last choice. On a percentage basis, this effect is greater for the larger good. Looking at the ADV treatment, one gets the impression that most of the reduction from the first to the last choice set is for the smaller good.[27] The first two rows of Table 7 provide *p*-values for the formal tests. These are marginally supportive of this notion and even more so if one were to use a one-sided hypothesis test.

We can compare point estimates for WTP at the same point in the sequence across treatments. The *p*-values in the last two rows of Table 7 imply no statistical significance in the difference across treatments given the same order in the sequence once the size of the good is accounted for. This is consistent with what one would expect under the notion that strategic behavior is taking place in both treatments along the sequence of choice tasks rather than the difference being completely incorporated into the first response. If there is a hint of an issue, it is with the small good and the first ADV response where the WTP estimates seem somewhat out of keeping with all of the other estimates in Table 6.

Table 6: P-Values for One Side Hypothesis Test of Null Hypothesis of No Difference Between Large and Small WTP Estimates (in £) Taken from Separate Probit Estimates

WTP Statistic	LARGE		SMALL		*p*-values	
WTP Statistic	median	mean	median	mean	median	mean
STP FIRST	13.07	17.90	4.74	5.56	0.004	0.015
	(1.78)	(4.08)	(2.52)	(3.51)		
STP LAST	8.88	11.66	3.90	5.14	0.038	0.117
	(1.82)	(3.56)	(2.00)	(3.48)		
ADV FIRST	11.71	12.84	7.89	9.62	0.197	0.308
	(3.35)	(4.29)	(2.47)	(3.93)		
ADV LAST	9.01	10.19	2.35	2.58	0.063	0.081

[27] In some ways, the last ADV response of suppressing WTP for the smaller good is similar to the Carson and Mitchell (2006) result using a sequence of open-ended questions whereby the respondents were able to revise the WTP amounts for reducing drinking water risks after seeing all of the risk levels.

Table 7: Empirical *p*-values for Differences Between Medians and Means (Two-sided Test)

	LARGE		SMALL	
	median	mean	median	mean
FIRST STP - LAST STP	0.053	0.128	0.396	0.475
FIRST ADV - LAST ADV	0.276	0.331	0.077	0.090
FIRST ADV - FIRST STP	0.643	0.811	0.190	0.227
LAST ADV - LAST STP	0.485	0.618	0.677	0.716

5.3 Parametric Analysis of Order Effects

We now move to a discussion of our analysis of ordering effects along the sequence of responses to all 17 choice tasks faced by our survey respondents. We begin by defining a linear multi-attribute utility function defined over the color, smell/taste and price attributes.[28] The linear indirect utility function specified in equation (3) is defined over all respondents and all choice tasks. In this equation, the vector of non-price attributes is specified as *x* and the water "Bill" variable is the cost to the respondent *n* of alternative *j* in choice task *t,* where we have information on the sequence *t=1,…,T* of responses collected for each respondent. We hypothesize that the marginal utility of income increases over the sequence of choice tasks. Given the specification of our utility function, the marginal utility of income is the negative of the coefficient on the price attribute variable. In order to directly account for the ordering effect on WTP, the *Bill* variable is interacted with the log of the question *order* (where the initial question has *order* = 1, the second question has *order* = 2 and so on):

$$U_{ijt} = \beta^{m} Bill_{ijt} + \beta^{order}(Bill_{ijt} \times \log(order)) + \sum \beta_i^k x_{ijt}^{\ k} + \varepsilon_{ijt} \qquad (3)$$

Ignoring the error component, the marginal utility of income (money) is given by:

$$\frac{\partial U_{ij}}{\partial Bill} = \beta^{m} + \beta^{l-ord} \log(order) \qquad (4)$$

A negative value for the coefficient β^{l-ord} means that the marginal utility of income increases over the sequence of choice tasks. Coupled with the log effect of order, this implies a gradual reduction of WTP estimates along the sequence with a more rapid decline taking place within the first few choice tasks. The marginal WTP for a generic attribute (*k*) is given by:

[28] The full factorial nature of our design would allow us to investigate much richer models but since the main focus of this paper is on testing the influence of ADV and STP treatments on the DCE format as typically implemented in the literature, the standard linear main effects model is adopted. A simple interaction term between color and odor is significant and negative, suggesting the two attributes are substitutes. However, inclusion of this term does not qualitatively change the results presented here.

$$WTP(order) = \frac{\beta^k}{-[\beta^m + \beta^{l\_ord} \log(order)]} \quad (5)$$

Estimates of the logit models are reported in Tables 8 and 9. The first two columns of Table 8 report estimated parameters for two conditional logit models: the standard multinomial logit (MNL) and a scaled version (MNL-scaled). The latter includes an estimate of the ratio of the scale parameter for those with advanced choice set awareness (ADV) to those with stepwise disclosure (STP). A ratio significantly larger than one implies that under the standard MNL specification, the error variance is smaller for those receiving the ADV treatment. This is what one would expect from notions of institutional learning (Braga and Starmer, 2005). It is noteworthy that all coefficients in both of these models are also significantly different from zero and consistent with *a priori* sign expectations.

Train (2003) argues that the MNL specification is predicated upon what are often implausible assumptions, specifically, the assumption of independence of irrelevant alternatives, fixed taste parameters and independence of choices by the same respondent. Panel logit models with continuous mixing of taste intensities are less restrictive. So, to explore the robustness of our results to the relaxation of such restrictive assumptions, we also report in Table 8 estimates from four different panel mixed logit models (MXLP). We assume that all mixing distributions for the random parameters are normal; however, the marginal utility of income is fixed to ease interpretation of the implied WTP distributions and to ensure identification. The first model (MXLP) assumes that the coefficients for status-quo (SQ_ASC), color and odor, are all distributed independently. Relative to the MNL models, the improvement in fit is quite large as shown by the (simulated) log-likelihood function values (LL). However, when we introduce a scaling factor to account for differences in ADV versus STP responses (MXLP-scaled), the estimated scale parameter is only a little larger than one and the difference is no longer statistically significant. Thus with preference heterogeneity accounted for, the residual variance of the common Gumbel error is not significantly different between the ADV and STP treatments.

The last two models relax the assumption of independence of preferences for two attributes, color and odor, by estimating the terms of the Choleski matrix related to the covariance between them. The first of these models (MXLP-corr) does not estimate a scale parameter to account for differential ADV versus STP responses, while the second (MXLP-corr-scale) does allow for such differences. In both cases the estimates of the Choleski matrix are significantly different from zero, thereby supporting the existence of significant correlation

effects. The ADV scale parameter is slightly larger than one, but this difference is not at all statistically significant. As such, the mixed logit models do not provide support for the notion of faster institutional or preference learning in the ADV versus STP treatments. This provides some weak support for the notion that the effect being observed is the result of strategic behavior rather than institutional or preference learning.

Table 8: Estimates of Conditional and Random Parameter Logit Models

	MNL	MNL-scaled	MXLP[a]	MXLP[a]-scaled	MXLP[a]-corr	MXLP[a]-corr-scale
Parameters	Estimates (|*z*-values|)	Estimates (|*z*-values|)	Estimates (|*z*-values|)	Estimates (|*z*-values|)	Estimates (|*z*-values|)	Estimates (|*z*-values|)
SQ_ASC (μ)	1.46 (28.6)	1.33 (20.1)	2.86 (17.3)	2.71 (12.5)	2.89 (17.9)	2.77 (12.3)
SQ_ASC (σ)			2.78 (17.6)	2.75 (13.4)	2.92 (20.4)[b]	2.81 (13.9)[b]
Bill	-0.0366 (15.5)	-0.0333 (13.7)	-0.0861 (15.9)	-0.0822 (12.5)	-0.0856 (16.1)	-0.0822 (12.2)
*Bill*Log_order*	-0.00815 (7.8)	-0.00746 (7.7)	-0.0178 (10.9)	-0.0172 (9.6)	-0.0177 (10.1)	-0.0171 (9.5)
Color (μ)	-0.296 (28.7)	-0.27 (20.3)	-0.064 (22.4)	-0.626 (13.9)	-0.656 (23.4)	-0.631 (14.0)
Color (σ)			0.41 (12.6)	0.367 (11.4)	0.381 (15.6)	0.368 (12.0)
Odor (μ)	-0.174 (32.3)	-0.158 (20.1)	-0.0367 (23.1)	-0.352 (14.8)	-0.368 (24.6)	-0.354 (15.0)
Odor (σ)			0.238 (12.6)	0.234 (10.0)	0.236 (13.3)	0.227 (10.5)
λ(ADV)/ λ(STP)=1		1.21 (2.65)[b]		1.08 (0.68)[b]		1.08 (0.69)[b]
C(color,odor)					-0.0393[c] (2.1)	-0.0381[c] (2.1)
LL	-6893.16	-6877.43	-4730.22	-4727.61	-4726.22	-4724.89
BIC	13821.28	13795.43	9521.29	9522.84	9520.06	9524.16

Notes:
[a] All mixed logit models are estimated with 350 Latin hypercube-draws in probability simulation.
[b] The *z*-value is computed with respect to the null hypothesis that the scale parameter equals one.
[c] Values of the Choleski matrix were used in the estimation to create random draws for standard normals.

Table 9: Estimates of Latent Class Mixed Logit Models

Parameters	2 Class LCM Estimates (|z-values|)		3 Class LCM Estimates (|z-values|)			4 Class LCM Estimates (|z-values|)			
	Class 1	Class 2	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 4
Membership Probability	.6135 (84.1)	.3865 (20.7)	0.407 (16.0)	0.467 (14.0)	0.126 (10.5)	0.273 (15.1)	0.346 (21.4)	0.077 (4.2)	0.304 (17.4)
SQ_ASQ	2.382 (17.0)	0.536 (10.2)	3.901 (11.3)	1.160 (19.9)	0.541 (4.8)	154.353 (0.0)	1.283 (12.4)	0.582 (12.4)	1.038 (15.9)
Bill	-0.078 (10.1)	-0.043 (18.4)	-0.079 (4.9)	-0.077 (28.6)	-0.012 (2.4)	-0.042 (2.0)	-0.127 (22.9)	-.9e-4 (0.0)	-0.053 (19.2)
*Bill*Log-order*	-0.020 (6.0)	-0.010 (9.9)	-0.027 (3.7)	-0.014 (11.3)	-0.013 (7.0)	-0.020 (1.8)	-0.026 (9.6)	-0.011 (4.1)	-0.015 (11.8)
Color	-0.439 (20.0)	-0.409 (37.4)	-0.597 (11.7)	-0.508 (40.7)	-0.486 (21.2)	-29.331 (0.0)	-0.554 (29.7)	-0.464 (14.8)	-0.536 (38.2)
Odor	-0.208 (-18.9)	-0.267 (44.6)	-0.218 (10.9)	-0.292 (43.5)	-0.394 (28.6)	-0.783 (2.6)	-0.285 (28.2)	-0.540 (23.1	-0.304 (41.3)
LL	-5398.252		-4920.53			-4736.51			
BIC	10870.88		9956.01			9628.54			

The significance of the coefficient for the price interacted with log-order is consistent throughout all models. Thus, our parametric results confirm the significance of an ordering effect such that respondents display a gradually higher marginal utility of income, and hence, lower marginal WTP as they progress through the sequence of choice tasks.

The log-transform is still significant in a heteroskedastic MNL model (available from the authors) in which 16 order-specific scale parameters were fitted. The point estimates for the (relative) scale are usually higher than one, indicating a reduction in the variance of the Gumbel error relative to the first response. Recall from Section 4.1 that the study was designed to test the difference in scale between the first and last questions which repeat the same choice. The estimated scale parameter for the last choice is 1.22, with the scale of the first choice normalized to 1. This is marginally significantly different from 1 (p=.09), and hence, suggestive of the usual notion of learning. However, the pattern of changes in the scale parameters across the whole sequence is somewhat erratic. Only for questions 3, 9, and 16 is the scale factor significantly different (at the .05 level) than 1 using the usual Swait and Louviere (1993) test. The least reliable estimates are in the middle of the sequence of questions. This position in a sequence is often where the most reliable estimates have been

previously found by the scale criteria. The most reliable question is the sixteenth one, a point by which fatigue is often said to have set in.[29] Further, making the standard Bonferroni-Sidak correction (Abdi, 2007) which is needed when undertaking multiple comparisons, shows that none of the individual scale parameters are significantly different from one. As such, we find at best weak evidence from this dataset to support a simple learning story cast in terms of an increase in the scale parameter.

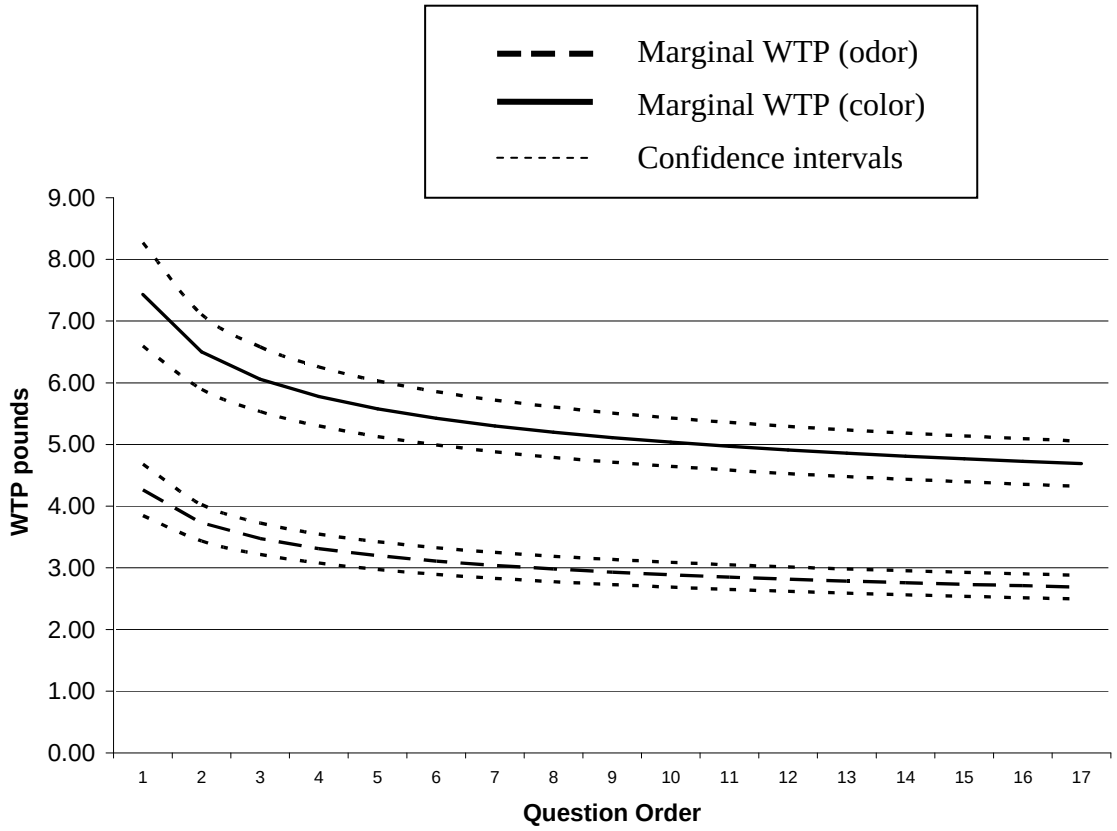
The other commonly used approach to deal with unobserved heterogeneity in a panel of discrete choice data is the latent class model.[30] Rather than require the researcher to make assumptions about the nature of the distributions for particular parameters, it assumes that the population of respondents is composed of a number of segments specified by the researcher. Each respondent has an estimated probability of membership for each segment. Preferences are homogeneous within a particular segment (*i.e.,* relative taste intensities assume the same value within each group) but heterogeneous between segments. There is no clear *a priori* rationale to prefer one representation of heterogeneity over the other (Provencher, Barenklau, and Bishop*,* 2002; Hensher and Greene, 2003; Scarpa, Willis and Acutt*,* 2005), nor are there powerful statistical tests that one can perform in order to select between the two approaches. In Table 9 we report estimates using the same utility function described above for a number of different latent class models (LCM): specifically we look at 2, 3, and 4 classes.[31] We do so in order to ascertain whether the conclusion drawn on previous results and support of the null of the existence of an ordering effect is sensitive to the form of heterogeneity used. As can be seen by inspecting the individual *z*-values, the coefficient on the interaction between price and log-order remains strongly significant across classes. The different parameter estimates for each class suggest that there are considerable differences between how the different latent classes respond to the price $\times$ log-order interaction, but the effect is negative and significant for all groups. Thus, our finding of increasing marginal utility of income (decreasing marginal WTP) is not dependent upon the method chosen to account for preference heterogeneity.

Figure 5 illustrates the implications of our findings of an increasing marginal utility of income (order effect) upon marginal WTP for both the discoloration and odor problems. These values vary systematically with the log of question order as implied by the estimates for the

[29] A quadratic function in terms of the effect of order on scale as might be suggested by a combination of learning and fatigue has the wrong signs but is not significant.
[30] A potentially useful approach not pursued here is the heteroskedastic logit (Louviere, Hensher and Swait, 2000).
[31] We also estimated a five class model but the BIC criteria indicated that such a model over-fits the data.

MXLP model in Table 8.[32] The magnitude of ordering effect is not only statistically significant but substantial in an absolute sense. However, it is interesting to note that most of the effect occurs within the first few response rounds. This seems intuitively plausible, as the realization of strategy space should set in quickly once the first question is followed by a second and so on.

Figure 5: Marginal WTP from MXLP Estimation Plotted against Question Order



[32] A generalization of the log-transformation was also tested using the Box-Cox transformation of order in the sequence; however, the transformation parameter was found not to be significantly different from zero so that the appropriateness of the log-transformation cannot be rejected.

6. CONCLUSION

The move from a SBC elicitation format to the multiple response approach of DCE studies brings with it many potential advantages in terms of the ability to ability to estimate the marginal value of changes in those attributes using a single reasonably sized sample. However, obtaining responses to multiple preference questions from a consumer also implies a loss of incentive compatibility. This may give rise to an array of possible strategic behaviors including cost-minimization and/or cost-averaging. While the incentive properties of the SBC format has been extensively researched, relatively little consideration has been given to how these effects play out when valuing a pure public good with the DCE method. This paper offers a contribution to that debate.

Awareness of the strategy and learning space afforded by DCE approaches can arise in two ways. First, respondents can be explicitly given such advanced awareness by receiving information before the DCE questions are asked about the attributes, levels and specific choice sets they will face (the ADV treatment). This contrasts with the standard stepwise (STP) approach wherein respondents have no prior awareness of attributes, levels or the choice set. Second, respondents can glean awareness of the choice set as they pass through the DCE exercise. This is the typical means by which STP respondents gain such awareness although such response experience may reinforce the awareness of ADV respondents. We term any effects arising from these two routes as choice set awareness and ordering effects, respectively.

The present study describes a design capable of testing for both choice set awareness and ordering effects. The design is constructed to allow both simple yet robust repeated measures ANOVA testing alongside more conventional parametric binary choice modeling. The results of both forms of testing are consistent: very clear and substantial choice set awareness effects are observed. Those facing the ADV and STP treatment give very different initial responses regarding the goods on offer. Analysis of the patterns of initial response observed suggests that a cost-averaging model provides the best description of the initial behavioral response by ADV respondents when compared to the initial STP response, which is assumed to be incentive compatible. This results in a considerable flattening of the bid response curve in the ADV case relative to the STP case. The ADV initial responses fail a scope test and there is little to suggest advanced disclosure helps improve the quality of the WTP estimates obtained.

Dynamic ordering effects are present in both the ADV and STP cases with the STP case being somewhat better behaved. These resemble what we have termed 'weak cost minimization' whereby at least some respondents say no to a choice that, taking the initial STP case as the true response, should increase utility. The incentive for this behavior is the potential ability to get the desired good for an even lower price. Strategic behavior is not the only interpretation of what may be occurring as respondents proceed through a DCE exercise. However, assessments of order-specific scale parameters fail to provide clear support for a learning hypothesis suggesting that this may be less important than strategic behavior, at least within the present DCE study.[33]

In conclusion, our findings suggest that within the context of valuing a public good, use of the DCE format can lead to significant advanced awareness and question order effects. These effects can translate into substantial impacts upon derived WTP estimates. Such effects deserve consideration within future applications and research.

[33] The general confounding of some types of strategic effects and scale along with the difficulties involved in developing designs appropriate for testing whether scale changes with sequence order noted in Section 4.1 are sure to keep this an active area of research.

References

Abdi, Herve (2007), “The Bonferroni and Sidak Corrections for Multiple Comparisons,” in *Encyclopedia of Measurement and Statistics*, Neil Salkind, ed. (Thousand Oaks, CA: Sage).

Adamowicz, W.L., J.J. Louviere, and M. Williams (1994), “Combining Revealed and Stated Preference for Valuing Environmental Amenities,” *Journal of Environmental Economics and Management*, 26: 271-292.

Adamowicz, W.L., P. Boxall, M. Williams and J.J. Louviere (1998), “Stated Preference Approaches for Measuring Passive Use Values: Choice Experiments and Contingent Valuation,” *American Journal of Agricultural Economics*, 80: 64-75.

Arrow, K., R. Solow, P.R. Portney, E.E. Leamer, R. Radner and H. Schuman (1993), Report of the NOAA Panel on Contingent Valuation, *Federal Register,* 58: 4601-4614.

Bateman, I.J., D. Burgess, W.G. Hutchinson and D.I. Matthews (2008), Contrasting NOAA Guidelines with Learning Design Contingent Valuation (LDCV): Preference Learning Versus Coherent Arbitrariness, *Journal of Environmental Economics and Management*, 55: 127–141.

Bateman, I.J., M. Cole, P. Cooper, S. Georgiou, D. Hadley and G.L. Poe (2004), “On Visible Choice Sets and Scope Sensitivity,” *Journal of Environmental Economics and Management*, 47: 71-93.

Bennett, J. and R. Blamey (2001), *The Choice Modeling Approach to Environmental Valuation*. (Cheltenham, UK: Edward Elgar).

Boyle, K.J., M. Morrison, and L.O. Taylor (2002), “Provision Rules and the Incentive Compatibility of Choice Surveys,” unpublished paper, Georgia State University.

Braga, J. and C. Starmer (2005), “Preference Anomalies, Preference Elicitation and the Discovered Preference Hypothesis, *Environmental and Resource Economics*, 32:55-89.

Bullock, C.H., D.A. Elston, and N.A. Chalmers (1998), “An Application of Economic Choice Experiments to a Traditional Land Use: Deer Hunting and Landscape Change in the Scottish Highlands,” *Journal of Environmental Management*, 52: 335-351.

Carlsson, F. and P. Martinsson (2001), “Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments? Application to the Valuation of the Environment,” *Journal of Environmental Economics and Management*, 41: 179-192.

Carson, R.T. and W.M. Hanemann (2005), “Contingent Valuation,” in *Handbook of Environmental Economics*, K.G. Maler and J. Vincent, eds. (Amsterdam: North-Holland).

Carson, R.T., N.E. Flores, and N.F. Meade (2001), "Contingent Valuation: Controversies and Evidence," *Environmental and Resource Economics,* 19: 173-210.

Carson, R.T. and T. Groves (2007) “Incentive and Informational Properties of Preference Questions,” *Environmental and Resource Economics*, 37: 181-210.

Carson, R.T., W.M. Hanemann, and D. Steinberg (1990), "A Discrete Choice Contingent Valuation Estimate of the Value of Kenai King Salmon," *Journal of Behavioral Economics*, 19: 53-68.

Carson, R.T. and R.C. Mitchell (2006), “Public Preferences Toward Risk: The Case of Trihalomethanes” in *Handbook of Contingent Valuation,* A. Alberini, D. Bjornstad, and J.R. Kahn, eds., (Northampton, MA: Edward Elgar).

Caussade, S., J. de D. Ortuzar, L. Rizzi, and D. Hensher (2005), "Assessing the Influence of Design Dimensions on Stated Choice Experiment Estimates. *Transportation Research B,* 39: 621–640.

Crowder, M.J. and D.J. Hand (1990), *Analysis of Repeated Measures*, (London: Chapman and Hall).

Dellaert, B.G., J.D. Brazell, and J.J. Louviere (1999), "The Effect of Attribute Variation on Consumer Choice Consistency," *Marketing Letters, 10*: 139-147.

DeShazo, J.R. and G. Fermo (2002), "Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency," *Journal of Environmental Economics and Management*, 44: 123-143.

Farrar, S. and M. Ryan (1999), "Response-Ordering Effects: A Methodological Issue in Conjoint Analysis," *Health Economics Letters*, 8: 75-79.

Foster, V. and S. Mourato (2003), "Elicitation Format and Sensitivity to Scope: Do Contingent Valuation and Choice Experiments Give the Same Results?," *Environmental and Resource Economics, 24*: 141-160.

Garrod, G. D., R. Scarpa, and K.G. Willis (2002), "Estimating The Benefits of Traffic Calming on Through Routes: A Choice Experiment Approach," *Journal of Transport Economics and Policy, 36*: 211-231.

Hanemann, W.M. and B. Kanninen (1999), "The Statistical Analysis of Discrete-Response CV Data," in *Valuing Environmental Preferences*," I. Bateman and G.K. Willis, eds., (New York: Oxford University Press).

Hanemann, W.M., B. Kanninen, and J.B. Loomis (1991), "Statistical Efficiency of Double Bounded Dichotomous Choice Contingent Valuation," *American Journal of Agricultural Economics*, 73: 1255-1266.

Hanley, N., R.E. Wright, and W.L. Adamowicz (1998), "Using Choice Experiments to Value the Environment: Design Issues, Current Experience and Future Prospects," *Environmental and Resource Economics*, 11: 413-428.

Hensher, D. and W.H. Greene (2003), "A Latent Class Model for Discrete Choice Analysis: Contrasts with Mixed Logit," *Transportation Research B*, 37: 681-696.

Holman, I.P., P.J. Loveland, R.J. Nicholls, S. Shackley, P.M. Berry, M.D.A. Rounsevell, E. Audsley, P.A. Harrison, and R. Wood (2002), "REGIS: Regional Climate Change Impact Response Studies in East Anglia and North West England," report to United Kingdom Department for Environment, Food and Rural Affairs.

Holmes, T. and K.J. Boyle (2005), "Dynamic Learning and Context-Dependence in Sequential, Attribute-Based Stated-Preference Valuation Questions," *Land Economics* 81: 114-126.

Hunter, P,R. (2002) *Climate Change and Waterborne and Vector-borne Disease*, School of Medicine, (Norwich, UK: University of East Anglia).

Johnson, R., and B. Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint Studies?," Technical Report (Sequim, WA. Sawtooth Software Inc).

Kabat P., R.E. Schulze, M.E. Hellmuth, and J.A. Veraart, eds. (2002), "Coping with Impacts of Climate Variability and Climate Change in Water Management: A Scoping Paper," DWC-Report no. DWCSSO-01(2002), International Secretariat of the Dialogue on Water and Climate, Wageningen, Netherlands.

Louviere, J.J. (2003), "Complex Statistical Choice Models: Are the Assumptions True, and If Not, What Are the Consequences?," Keynote Address, Discrete Choice Workshop in Health Economics, University of Oxford.

Louviere, J.J. and D. Hensher (1982), “On the Design and Analysis of Simulated Choice or Allocation Experiments in Travel Choice Modelling,” *Transportation Research Record* 890: 11-17.

Louviere, J.J., D. Hensher, and J. Swait (2000), *Stated Choice Methods: Analysis and Application*, (Cambridge: Cambridge University Press).

Louviere, J.J., D. Street, R.T. Carson, A. Ainslie, T.A. Cameron, J.R. DeShazo, D. Hensher, R. Kohn, and T. Marley (2002), “Dissecting the Random Component,” *Marketing Letters*, 3: 177–193.

MORI (2002), *The 2004 Periodic Review: Research into Customers’ Views,* (London: DEFRA).

National Statistics (2001), *Regional Trends 2001,* (London: Office for National Statistics).

Ohler, T., A. Li, J.J. Louviere, and J. Swait (2000), “Attribute Range Effects in Binary Response Tasks,” *Marketing Letters*, 11: 249-260.

Phillips, K.A., F.R. Johnson, and T. Maddala (2002), “Measuring What People Value: A Comparison of ‘Attitude’ and ‘Preference’ Surveys,” *Health Services Research,* 37: 1659-1679.

Plott, C.R. (1996), “Rational Individual Behavior in Markets and Social Choice Processes: The Discovered Preference Hypothesis,” in *Rational Foundations of Economic Behavior*, K. Arrow, E. Colombatto, M. Perleman, and C. Schmidt, eds., (London: Macmillan).

Poe, G.L., M.P. Welsh, and P.A. Champ (1997), “Measuring the Difference in Mean WTP When Dichotomous Choice Contingent Valuation Responses Are Not Independent,” *Land Economics*, 73:255-67.

Provencher, B., K. Barenklau, and R. Bishop (2002), “A Finite Mixture Model of Recreational Angling with Serially Correlated Random Utility,” *American Journal of Agricultural Economics,* 84: 1066-1075.

Samuelson, P.A. (1954), “The Pure Theory of Public Expenditure,” *Review of Economics and Statistics,* 36: 387-389.

Scarpa, R., E.S.K. Ruto, P. Kristjanson, M. Radeny, A. Drucker, and J.E.O. Rege. (2003), “Valuing Indigenous Cattle Breeds in Kenya: An Empirical Comparison of Stated and Revealed Preference Value Estimates,” *Ecological Economics*, 45: 409-426.

Scarpa, R. and K.G. Willis (2006), “Distribution of WTP for Speed Reduction with Non-Positive Bidders: Is Choice Modeling Consistent with Contingent Valuation?,” *Transport Reviews*, 26: 451-469.

Scarpa, R., K.G. Willis, and M. Acutt (2005), “Individual-Specific Welfare Measures for Public Goods: A Latent Class Approach to Residential Customers of Yorkshire Water,” in *Econometrics Informing Natural Resource Management*, P. Koundouri, ed. (Cheltenham, UK: Edward Elgar Publisher).

Street, A.P. and D.J. (1986), *Combinatorics of Experimental Design* (Oxford: Oxford University Press).

Swait, J. and W.L. Adamowicz (2001), “The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Switching Strategy,” *Journal of Consumer Research,* 28: 135-148.

Swait, J. and J.J. Louviere (1993), “The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models,” *Journal of Marketing Research,* 30: 305-314.

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

Tuan, T.H. and S. Navrud (2007), "Valuing Cultural Heritage in Developing Countries: Comparing and Pooling Contingent Valuation and Choice Modelling Estimates," *Environmental and Resource Economics,* 38:51-69.