Preference stability in discrete choice experiments. Some evidence using eye-tracking

Article

Accepted Version

Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

Fraser, I., Balcombe, K., Williams, L. and McSorley, E. (2021) Preference stability in discrete choice experiments. Some evidence using eye-tracking. Journal of Behavioral and Experimental Economics, 94. 101753. ISSN 2214-8043 doi: https://doi.org/10.1016/j.socec.2021.101753 Available at https://centaur.reading.ac.uk/99621/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1016/j.socec.2021.101753

Publisher: Elsevier

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading’s research outputs online

Preference Stability in Discrete Choice Experiments. Some Evidence Using Eye-tracking.

Kelvin Balcombe
School of Agriculture, Policy and Development
University of Reading
Reading
UK
k.g.balcombe@reading.ac.uk
Iain Fraser (Corresponding author)
School of Economics
University of Kent
Canterbury
UK
i.m.frasder@kent.ac.uk
Louis Williams
Dynamic Planning
Reading
UK
louiswilliams10kn@hotmail.co.uk
Eugene McSorley
School of Psychology and Clinical Language Sciences
University of Reading
Reading
UK
e.mcsorley@reading.ac.uk

Accepted for publication in Journal of Behavioural and Experimental Economics

July 2021

Preference Stability in Discrete Choice Experiments. Some Evidence Using Eye-tracking.

Abstract

We investigate the relationship between the extent of visual attention and preference stability in a discrete choice experiment using eye-tracking to investigate country of origin information for meat in the UK. By preference stability, we mean the extent to which choice task responses differ for an identical set of tasks for an individual. Our results reveal that the degree of visual attention, counter to our initial expectations, is positively related to the degree of preference instability. This means that preference instability does not necessarily indicate low levels of respondent engagement. We also find that those respondents' exhibiting preference instability do not substantively differ from the rest of the sample in terms of their underlying preferences. Rather, these respondents spend longer looking at tasks that are similar in terms of utility, suggesting these respondents find these choices more difficult.

Key Words: Discrete Choice Experiment, Eye-Tracking, Preference Stability, Complexity
JEL: D83, Q18, C99

1. Introduction

There is a growing body of literature that has investigated the link between preferences revealed by discrete choice experiments (DCE) and visual attention (number of fixation or dwell time) using eye-tracking (ET) (e.g., Balcombe et al., 2015, 2017; Van Loo et al., 2018). The motivation for these inquiries are several. First, it is proposed that ET data may be used to improve estimates of utility and willingness to pay (WTP) (e.g. Yegoryan et al., 2020). Second, ET data might be employed to improve the design of DCEs at the pilot stage prior to full implementation. Third, and the focus of the research presented in this paper, is the potential relationship between ET data, specifically visual attention and preference stability. We define preference stability in the context of a DCE as follows. If we ask a survey respondent to select an option from a choice set, preference stability means that they would consistently select the same option if shown the same choice set repeatedly. To examine preference stability, we provide a group of respondents a set of choice tasks and then without explicitly stating, we presented with the same set of choice tasks again. By examining the extent to which individual specific responses differ across the identical set of tasks, we can assess the level of preference stability.

Preference stability is an important issue for DCE research as it goes to the heart of DCEs which purport to rely on random utility model foundations, because preference stability is assumed. Furthermore, it is frequently assumed (explicitly or implicitly) in much preference based research underpinned by the random utility model, that respondents can undertake the necessary cognitive tasks such that all choices are "optimal", in that they maximise utility subject to constraints, which might be for example the level of complexity involved (Hess et al. 2018). Indeed, there is significant evidence demonstrating that choice complexity has a negative impact on the accuracy and precision of choices made (Swait and Adamowicz, 2001; Meißner et al., 2016; Rigby et al., 2016; Hess et al., 2018). By examining preference stability, we provide insights into the issue of choice complexity as it applies to DCEs. In particular, by using ET we are able to examine the extent to which survey respondents are engaging with the DCE. This in turn allows us to better understand the reasons why preferences may not be stable and what this implies for DCE design.

Our study specifically contributes to the literatures that have employed repeated choice tasks within stated preference studies in order to examine preference stability. Preference stability can be considered a specific form of preference reversal. The study of preference reversals has been subject of much research going back several decades e.g., Lichtenstein and Slovic (1971) and Grether and Plott (1979). In this literature reversals are examined, for example, by choice consistency between product type, choice context and survey methodology. In this study, we employ an approach similar to that used by, for example, Carlsson et al. (2012) and Segovia and Palma (2020), by repeating a set of choice tasks within a DCE using ET. Specifically, we present a group of survey participants with 24 choice tasks of which the first 12 are repeated. Like Segovia and Palma (2020), we have implemented our DCE using ET. However, unlike Segovia and Palma (2020), we examine choice behaviour not only on a task by task basis but also, we consider how much each specific attribute influences the choices made. This means, that we can examine the time taken to consider each attribute within the DCE. In addition, by measuring total dwell time for each choice task, that is the time taken to consider a choice card and make a selection, we can see how this relates to preference stability. Our research also examines the relative difference in speed of response to specific choice tasks and how this relates to choice task complexity. Our research contributes to the extensive body of research on this topic that needs to be considered when designing and evaluating DCE performance (Swait and Adamowicz, 2001; Balcombe and Fraser, 2011; Olsen et al., 2011; Pfeiffer et al., 2014; Regier et al., 2014; Meißner et al., 2016).

Our DCE was specifically designed to investigate country of origin (CoO) information for meat in the UK. The motivation for the study stems from EU proposals in April 2015 to extend the scope of mandatory CoO labels to include fresh, chilled and frozen meat of swine, sheep, goats and poultry as well as a further proposed extension to processed food products containing meat such as pies and pizzas.[1]

In general, we find that the majority of survey participants in this study pay reasonable attention to the tasks they are presented (albeit, they know that there eye-movements are being tracked). However, there is still a wide distribution of dwell time among participants by attribute and by tasks. It is possible that participants dwell and fix on choice task information for long periods without processing the information. However, we believe this is unlikely because dwell time measurements are for specific regions of the choice task where information is being presented. In fact, from the repeated choice task there is a strong *prima facie* case that

[1] The specific legislation is the European Commission Implementing Regulation (EU) No 1337/2013 of the 13th December 2013 laying down rules for the application of Regulation (EU) No 1169/2011 of the European Parliament and of the Council as regards the indication of the country of origin or place of provenance for fresh, chilled and frozen meat of swine, sheep, goats and poultry sets out the requirements for CoO for these species. These changes to CoO labels became mandatory for these products from the 1st April 2015 and was part of wider push for the introduction of mandatory labels on many other food types in the EU.

if respondents are fixing upon this information specifically, it is because it is being used to make a decision, and the reason why participants spend longer looking at specific tasks is because the choice alternatives have similar utility as measured by the "entropy" in choice probabilities. So the choice is more "complex".

Furthermore, our results clearly indicate that more complex choices not only take longer to make but also lead to more stochastic responses and preference instability. Put simply, those respondents who were presented with the repeated choice task and who had higher rates of visual dwell tended to exhibit higher levels of preference instability. Thus, higher rates of visual dwell are being driven by respondents visually rechecking information when a choice is not immediately clear. This is an important finding as it implies the lack of preference stability is not a function of respondent engagement with the DCE but rather the complexity of the task in hand. This conclusion provides support to those of Campbell et al. (2018) who observed that longer response times correlated with higher levels of measurement error within econometric models. In addition, we also find that the duration of dwell time declines moderately throughout the DCE. This decline in dwell time is frequently attributed to learning effects during the DCE and the degree of decline appears somewhat less than levels previously reported in the literature (Meißner et al., 2020).

Finally, Meißner et al. (2016) report that attributes with greater importance to respondents generate greater attention in this case measured as the number of fixations. However, we observe that salience of an attribute can mean that less attention (as measured by dwell time) is required even if the utility associated with an attribute is higher. Thus, although there is a tendency in the literature to report fixations or dwell time and support this choice by indicating that the measures are generally highly positively correlated, this is not always the case and the differences observed can have important behavioural implications.

The paper proceeds by first examining the antecedent literature in Section 2. Next, we describe the DE that examines food choice with a specific focus on CoO information for meat on a food label. We also provide extensive details of how we implemented repeated choice tasks plus our ET setup. We then provide a description of our econometric specification and approach to model estimation in Section 4. The results of our DE are presented in Section 5, and Section 6 concludes.

2. Antecedent Literature

2.1. Repeated Choice Tasks

A key feature of the research present in this paper is the fact that we employ a set of repeated choice tasks. Within the literature there are a number of related but subtly different approaches to how the choice tasks have been repeated. For example, there is the test-retest approach (e.g., Green and Srinivasan, 1978) that has been used in a number of DCE (e.g., Morkbak and Olsen, 2014; Rigby et al., 2016). In these studies, survey respondents complete a series of choices twice with varying degrees of time between attempts (e.g., minutes, hours, days). A potential weakness of this type of test-retest approach noted by Rigby et al. (2016) relates to the extent to which respondents self-select to participate in the retest part of a test-retest study. As they note the response rate can be significantly less than 100% and as such the composition of the retest sample may be biased (as a result of self-selection) towards respondents who make more consistent choices. Another approach is to repeat one or two choice task within the same survey. There are several examples of studies that repeat a single choice task (e.g., Johnson and Matthews, 2001; Mattmann et al., 2019). However, by only employing a single choice tasks to assess preference stability, the inherent complexity of the specific task used can influence the degree of preference stability observed. An alternative approach is to include a much larger number of repeated tasks. For example, Carlsson et al. (2012) required survey participants to complete 16 choice tasks, where the first eight were repeated. They report that up to 27% of respondents change their choice for the first choice with between 11 and 20% for subsequent tasks. More recently, Segovia and Palma (2020) conducted a DCE in which respondents undertook three versions of the same DCE with ET employed so as to examine issues of choice consistency. Their experimental design involved making minor modifications to the DCE so that they could examine how variation in spatial location of choice task attributes and choice options affect visual attention, search dynamics and valuations. They report that WTP estimates are generally consistent across all three versions in terms of with only minimal changes in the estimates presented. Other approaches reported in the literature to assess preference stability and consistency include simple tests using the underlying axioms of choice such as transitivity (e.g., Sælensminde, 2002) and comparisons of choice methodologies (e.g., Chen et al., 2020). Given our interest in preference stability and task complexity, we adopted an approach similar to that of Carlsson et al. (2012) and Segovia and Palma (2020), except not only do we employ ET but we also examine attribute specific features of the choices being made.

2.2. Task Response Times

As we note in the Introduction, we employed ET as part of the implementation of the DCE meaning that we can also examine response times, not only for the whole survey, but also by choice task and time spent considering each specific attribute. The fact that we are able to collect response time data means that our research adds to an extensive economic and decision theory literatures employing response times in various types of analysis (e.g., Konovalov and Krajbich, 2017; Clithero, 2018; Spiliopoulos and Ortmann, 2018). For example, a specific use of response time data is in sequential sampling models such as the Drift-Diffusion Model (DDM). In this context response time data are used to examine choice processes (Krajbich and Rangel, 2011). Although the use of DDMs is wide ranging there application to DCE data in economics has yet to occur.

There is already significant interest in response times within the DCE literature. For example, as noted by Uggeldahl et al. (2016) who undertook a DCE using ET, it has been suggested that low levels of dwell time (i.e., quick response times) could be indicative of low respondent engagement, and by consequence higher rates of preference instability. Other DCE studies that have examined response time, although not using ET, include Börger (2016) and Campbell et al. (2018). In these studies, response times are typically recorded via the mouse on a click-by-click basis over the entire duration required to complete an individual choice task and or set of choice tasks. Response times are then used to examine the error variance of the econometric models estimated. The rationale for this is explained by Hess et al. (2010) who note that trying to identify choice inconsistency directly is difficult but the impact will manifest via the stability of individual coefficients. Interestingly, Campbell et al. (2018) find that the level of measurement error increased as response times increased. Our analysis is able to contribute to this research question. In particular, by employing ET not only do we obtain a measure of how long it takes to complete a specific choice task, we can also examine what aspect of the choice task is completed at different speeds. This insight is particularly important as it helps us to understand how respondents engage, that is taking more or less time, with the information provided and the task in hand.

2.3.Choice Task Complexity

As explained by Pfeiffer et al. (2014), there are significant differences in how complexity has been defined in the choice literature that impact how it is examined. First, there is task-based complexity that describes general features of a choice task such the number of alternatives and/or the number of attributes. It can also include how the choice task is presented. Second, there is context based complexity that relates to the individual undertaking the choice task. In this case complexity encompasses the difficultly of the task to the individual respondent given specific attributes, attribute levels and the similarity of alternatives. In this paper, the complexity we are concerned about relates to the level of effort required to make a choice.

This distinction in terms of the specific meaning of complexity is, however, less apparent in many of the empirical studies that have examined complexity. There are various empirical measures of complexity that have been used within the literature to date.

First, a common way of describing complexity within a DCE is simply to assess the quantity and type of information presented to survey respondents. This type of complexity can include the number of alternatives within a choice set as well as the number of attribute and the associated number of levels. For example, many DCE studies define design complexity in terms of number of choice tasks (e.g., Bech et al. 2011; Carlsson and Martinsson 2008; Hensher et al., 2001) or the number of experimentally designed alternatives, attributes, and attribute levels (e.g., DeShazo and Fermo 2002; Arentze et al., 2003; Hensher 2006a,b; Caputo et al., 2017).

In the context of food choice it has been argued that consumers use food attribute information in different ways depending on the type of attribute. In particular, a distinction is drawn between cue versus independent attributes, as defined by Gao and Schroeder (2009), and how the combination of attribute types might impact on task complexity. CoO is an example of cue attribute in that it not only conveys a specific meaning but it can also implicitly provide information about other product attributes not used in the DCE. In contrast an independent attributes has a clear and very specific meaning. Understanding this distinction, researchers have noted that as complexity of a DCE increase, with the inclusion of more independent attributes, we can observe increased effects in terms of learning and/or fatigue (Caputo et al., 2017).

In practice, many of the experimental design features that generate complexity are fixed at the design stage. However, how information is subsequently presented within a choice task can also influence complexity. This type of complexity can relate to the number of trade-offs required to make a choice or the degree of attribute value dispersion which can be measured in a number of ways (Pfeiffer et al., 2014). To date several studies employing ET have examined how design features impact choice complexity. For example, Meißner et al. (2016) present DCE with all attributes in separate rows which can in principle help an attribute-wise search processes. They also used four options within each choice task which might give rise to stronger alternative-focus and attribute-focus effects. Indeed, in subsequent research, Meißner et al. (2020) note that the number of alternatives in a task influences the strength of the alternative-focus and attribute-focus effects and how much they change in later tasks.

Second, there is the over-arching view of complexity introduced by Swait and Adamowicz (2001) who employed entropy as a measure of complexity. In this context entropy is defined in the standard way such that it refers to a measure of the uncertainty around a random variable. The higher the level of entropy is taken as evidence of greater complexity within the context of a DCE. The use of entropy to describe complexity has proven very useful within the choice literature. In general, as complexity increases respondents will need to employ higher levels of cognitive effort, but there comes a point at which the complexity of the task is such that respondents start to employ choice heuristics and not use all of the information available, because of the complexity, and as such start to make inconsistent choices. Thus, when presented with simple tasks respondents will be able to identify the preferred option but as complexity increases the likelihood of inconsistent choices will increase. Thus, it is when complexity in a choice set is at a medium level that we are likely to obtain the highest level of variation in responses as some respondents will be able to cope with task in hand whereas others will resort to using some sort of simplifying heuristic. In this paper, we follow Swait and Adamowicz (2001) and employ an entropy based measure that is derived *ex-post* from the relative utility derived from the choices made.

As noted several studies that have examined complexity using ET. For example, Pfeiffer et al. (2014), showed that complexity significantly influences information acquisition, such that information search during a choice task is more 'attribute-wise' as opposed to 'choice-alternative-wise'. They also report that as a choice task becomes more complex it generally takes longer, but the salience of attributes did not have an impact on the order in which attributes were attended. In another conjoint study using ET, Meißner et al. (2016) present results from three studies. In each study they use the number of fixations as their measure of attention although they report that similar results are found if employing time spent considering each attribute i.e., dwelling. In general, they find that respondent attention during a set of choice tasks is directed toward alternatives yielding higher levels of utility and as well as attributes yielding higher utility. Our analysis will contribute to the general body of literature that has considered task complexity using ET

3. The DCE

3.1. Attributes and Design

The ET DCE survey data that we employ in this paper, was collected to examine how respondents engaged with a DCE survey instrument. This data set is part of a bigger study undertaken to examine the value UK consumers place on CoO for meat products (see Balcombe et al., 2016, 2017). For example, Balcombe et al. (2016) considered 12 meat based food products. These studies also employed DCE survey instruments that used fewer attributes (five as opposed to seven). Thus, in this study, we use two additional attributes, the number of calories per pizza and the country of production of the pizza. These attributes allowed us to examine if a source of health information impacts choice and if the country in which the pizza is produced is important, as opposed to the country from which the meat is sourced matters to respondents.

The specific reason for including both country of product production and the CoO of the meat used within the pizza was to see if respondents placed more importance in one attribute over the other. Also, the distinction between country of product production and CoO of specific ingredients has not been examined in detail for processed products within the literature, and it is potentially an important issue to consider if country's wish to employ mandatory CoO for meat in processed products.

The rationale for the design of this specific DCE was to examine survey participant engagement with the survey instrument, the relationship between ET data and WTP, and preference stability and complexity. The product we used to examine CoO is a standard supermarket bought pepperoni pizza. The set of attributes used to describe the pizza are as follows:

- **Price (PR)** - For this attribute, the range of values was determined by reference to product size and description and by reference to those most commonly on sale in UK shopping outlets. The set of prices used in the DCE are £2.00, £2.95, £3.75 and £5.25;
- **Country of Origin (CoO)** - This attribute indicates the origin of the meat and had four possible options: UK, USA, Italy and EU. The choice of countries (USA and Italy) reflects potential sources for imports of pepperoni and the UK and EU capture home country and a generic source indication that is used on existing food products;
- **Product Quality (PQU)** - We selected three levels for this attribute: Basic, Choice and Premium. The inclusion of an attribute to describe product quality meant that we could implicitly capture aspects of the product that relate to taste or other quality related characteristics. This attribute acts a cue indicating all of those characteristics that constitute "quality", but are not explicitly stated;
- **Farming System (ORG)** - This attribute was either Organic or Conventional. These two production systems capture the majority of meat production on the market and is familiar to consumers;
- **Quality Assurance (QAS)** – For this attribute our levels are: No label, Freedom Food and the International Quality mark;
- **Calories (CAL)** - with levels 195, 255,315, 350, 524, 714 (per 100 grams); and,
- **Country of Production (CoP)** - This is the country where the Pizza was produced.(UK vs EU).

Full details of the experimental protocol we employed in the study are given Appendix E. The material presented in Appendix E explains how the collection of the DCE was framed as well as details of the additional information collected as part of the study.

Given the above set of attributes, an example of a choice card used in the DCE is presented in Figure 1.

[Approximate Position of Figure 1]

One feature of the choice card presented in Figure 1, is that for some of the attributes there are blank cells. This occurs because when, for example, we consider a specific product it may or may not uses organically produced ingredients. Although there is an accepted label to signify organic there is no need to label a product as not being organic. This design issue has previously been noted by Uggeldahl et al. (2016) who note that using a non-standard label to fill a blank may, in fact, not be the best way to replicate a real-world shopping experience. The implications of designing our choice cards in this manner is something that we need to be cognizant of when it comes to analysing our data.

Given the choice of attributes and associated levels, we designed our choice sets in a standard manner. We employed a conservative efficient design assuming a Multinomial Logit utility specification employing D-error as the measure of design efficiency. Our design was produced using Ngene version 1.1.1 (Choice Metrics 2012) assuming null priors on our model coefficients. In total, we generated 24 choice cards for the DCE with two product profiles on each. All of the survey participants completed the 24 choice cards which where randomised by task order.

An important feature of this specific DCE is that half of the respondents were randomly selected and given 12 cards twice (randomly selected from the original 24). The rationale for approach was to allow us to consider the extent to which choices were made consistently. That is, we wished to consider the extent of choice selection for the same card and if the selected option changed.

Although 24 choice cards is more than typically used in a DCE, we consider it is appropriate and not excessive. Indeed, there are plenty of existing studies that employ a high number of choice tasks and this is especially the case when examining consumer choice with regard to goods and services that are well understood by respondents (e.g.,Hensher et al. 2001; Louviere 2004; Carlsson and Martinsson 2008; Bech et al. 2011;). Indeed, Louviere (2004) goes as far as stating:

"...It is widely believed that 'modeling' individuals requires 'smallish designs,' but in contrast to the equivalent of widely held 'academic urban myths' in marketing and transport research, there is considerable evidence that humans will 'do' dozens (even hundreds) of T's (choice tasks)"(p. 18).

Furthermore, given the objective of the research presented there needs to be many choice sets to meaningfully explore preference stability. Repeating only four cards for example would not give enough information and only selecting four cards could bias the relative complexity of the repeated tasks. Moreover, the current "convention" of giving respondents 8-12 cards as a maximum is little more than a convention based on the belief that more than this number of choice tasks will induce respondent fatigue. In keeping with standard practice our survey instrument also provided all participants with an overview of the survey context prior to undertaking the DCE.

As is standard within DCE, we employed a set standard de-briefing questions after all the choice cards had been completed. The data was collected using a standard question format: *"Which of the following attributes did you ignore when completing the choice task? (You can tick none or as many as required)"*. This data allowed us to construct stated attribute non-attendance (SANA) variables for all of the DCE attributes.

In designing and implementing this specific DCE, we did not include a no-choice option. We made this choice because the DCE was implemented with a focus on survey participant engagement with the choice tasks. The absence of a no-choice option is quite common in DCE and conjoint studies that employ ET data e.g., Meißner et al. (2016, 2020). The authors did not include an opt-out option. Although they motivated the non-use of an opt-out option, In addition, we note that Boxall et al. (2009) in a study on the influence of complexity argued that excluding the status quo option, which frequently plays a similar role to the no-choice option in DCE, allowed them to focus in detail on the experimentally designed attributes. Finally, given that the sample of respondents we employed are not a representative sample of consumers, we make no attempt to extrapolate the meaning of our resulting WTP estimates to the wider population.[2]

3.2. Survey Participants

We implemented the DCE with 100 participants although one respondent was dropped from the final sample. Although not large by typical non-ET DCE standards our sample is comparable with many other ET DCE studies in the literature (e.g., Krucien et al. (2017), n= 58). However, given the relatively modest size of the sample we employ the results generated need to be treated with a degree of caution. The sample was recruited at the University of Reading in the UK via email with a £10 participation fee offered. The final sample was composed of 53 females and 46 males. The sample contained a wide range of ages, but with a larger proportion of young people than in the UK population as a whole with very few participants over 55 years of age. As such the average age of the sample of participants was 31 years old.

All survey participants indicated that they consumed meat and we confirmed this by employing a screening question at the start of the survey. In addition, almost all (96) indicated that they were either the main shopper in the household (60) or shopped for meat some of the time (36). The majority of these participants indicated that they bought fresh meat more commonly than frozen, usually at least once a week and that they shopped in the expected range of supermarkets.

3.3. ET Implementation

3.3.1. Apparatus and Room Setting

To implement our ET choice cards were presented on a 21 inch colour monitor with a refresh rate of 75 Hz. The viewing distance was 57 centimeters (cm) such that cm are equal to degrees of visual angle. Eye movements were recorded using a head-mounted, video-based, eye-tracker with a sampling rate of 500 Hz (Eyelink II, SR Research), recording monocularly from the respondents' right eye. Head movements were constrained with a chin-rest, which held the

[2] The DCE presented here was part of a large online DCE examining CoO for 12 different meat products using an almost identical form of choice card. The only difference was that, first, we forced a choice between two options and then offered respondents the choice to keep the product selected or to select no choice. This type of question design is referred to in the literature as dual-response method (Brazell et al., 2006). The analysis of this data accounting for the no-choice responses yielded results almost identical to those for the forced choice. Although not directly comparable to the current study these previous results suggest that the exclusion of a no-choice option was a reasonable design decision.

participant so that their eyes were in line with the horizontal meridian of the screen. Choices were recorded through a response gamepad. The room in which the ET was undertaken was completely dark. The room was also enclosed so that nobody could walk in while testing was carried out and disturb participants.

The eye-tracker was calibrated using a standard 9 point grid, carried out at the beginning of the DCE. The background colour of the calibration was white (255,255,255). Calibration was accepted only once there was an overall difference of less than 0.5 degrees between the initial calibration and a validation retest. In the event of a failure to validate, calibration was repeated. In order to ensure that accuracy was maintained throughout the DCE a drift correction was carried out between each card viewing. Participants were asked to view a spot stimulus and press a button when they were fixating its centre. The drift correction stimulus consisted of a small black annulus that gave the appearance of a small black spot (0.5 centimeters in diameter) with a smaller white spot in the centre (0.25 cm diameter) shown in the upper left quadrant of the screen off set from the centre by 5.12 degree horizontally and 3.86 degree vertically. This procedure minimized the effects of slight movement of the head impacting on the accuracy of the eye-tracking.

Once participants were comfortable in the eye-tracker and their eye movements calibrated, they were presented with the series of choice cards. Participants viewed the choice cards for as long as they wished while we tracked their eye movements. They responded with a button press indicating which product they selected. A drift correct stimulus was then shown until a button press from the participant indicated they were looking at it. The next choice card was then shown.

3.3.2. Choice Cards and Areas of Interest (AOIs)

In designing the DCE for the ET version of the survey all choice cards were presented on a white background. Although there could be minor differences between choice cards in terms of luminance it was not influential on the results. Furthermore, as already noted the viewing distance (eyes to screen) was 57 cm. Given this viewing distance it then follows given the design of the choice cards that each choice option was 12.4 by 13.8 cm (or degrees) in size on the screen.

In Figure 2, we show the respective AOIs for the DCE as yellow boxes.

[Approximate Position of Figure 2]

The AOIs displayed in Figure 2 are shown with size areas. These areas are 4.5 by 3.6 cm (degrees) around each of the text boxes (e.g., Meat Origin, etc.) and 8.9 and 2.8 am (degrees) around the top part of each card (i.e., Option A). Given the guidance in Orquin and Holmqvist (2018) it is stated that a 3.2 degree AOI will yield an 80% capture rate. Although, we do not employ circular AOIs due to the nature of the stimuli, our cards and the AOIs are bigger than the guidance provided. Therefore, we can be confident that the stimuli are of sufficient size to carry out our analysis.

To ensure that our ET results are robust, we employed a tight version of the AOI, centred on each rectangle and one which was more generous (1 or 0.5 degree around the outside). Relaxing the AOI made no difference to the ET results generated. Thus, as we make clear, our approach to defining the AOI was such that we have not modified them *ex-post* the ET data collection and as such cannot be subject to the criticisms of such ad hoc practices discussed by Orquin et al. (2016). In terms of the data collected only 0.19% of trials were not included in the analysis due to missing data. Also, 18.1% of fixations were excluded due to not falling within the created AOIs. All fixations greater than 100ms and within the AOIs were included in the analysis. Finally, we did not randomise the location of the attributes during the choice task but we did randomised the order of the choice tasks.

3.3.3. Data Collected and Definitions

Although there has been a rapid growth in the use of ET within economics as well as a much longer use of the technology in other disciplines it is always essential to be clear about how ET data is being interpreted. Here we follow the definitions introduce by Balcombe et al. (2015, 2017) who distinguished between visual attendance and attention:

- **Visual attendance (attended)**: this requires a respondent to "fix" on an attribute for all choice options for a choice task. Requiring all options are fixed upon (unless blank) is needed for a fair comparison of attribute levels to have been made.
- **Visual attention (dwell time):** this is measured by the total "dwell time" on a particular attribute (i.e., how long looked at).

The ET data we will generally employ in the subsequent analysis is dwell time by attribute and choice set. However, we will also employ the measure of visual attendance in the resulting analysis as this allows us to explore specific aspects of choice behaviour.

4. Model Specification and Estimation

This study uses a generalisation of the 'mixed logit' for estimation because it can approximate a wide range of random utility models, and allows for respondent heterogeneity when making choices. The mixed logit can be implemented within a classical or Bayesian statistical framework, and within the latter framework the mixed logit is commonly referred to as the Hierarchical Bayes Logit (HBL). As is standard in the DCE literature, we assume that the utility (U_{ijs}) for the *jth* person from the *ith* option in the *sth* choice set is:

$$U_{ijs} = u\left(x_{ijs}, z_j\right) + e_{ijs} \tag{1}$$

where x_{ijs} is a $(K \times 1)$ vector of known attributes and z_j is a vector of observed characteristics for the *jth* respondent. We also assume an extreme value error e_{ijs} that is independent across, i, j and s implying that the probability of choosing option i for the *jth* person from the *sth* choice set is:

$$p_{ijs} = \frac{e^{u(x_{ijs}, z_j)}}{\sum_i e^{u(x_{ijs}, z_j)}} \tag{2}$$

The form of the utility function specified in this paper is:

$$u\left(x_{ijs}, z_j\right) = \sum_{k=1}^{K} \beta_{jk} x_{k,ijs} \tag{3}$$

where β_j is a $(K \times 1)$ vector that is a function of z_j which is vector specified as $z_j = (\delta_{j1}, ..., \delta_{jK})$, such that $\delta_{jk} = 1$ if the jth individual states that they ignore the kth attribute and zero otherwise.

As noted in Section 2, in this DCE, we employed a standard de-briefing question after the choice task is completed. This question allows us to generate stated attribute non-attendance (SANA) data.[3] This data is integrated into the model by defining:

$$\rho_{jk} = \left(1 - \frac{1}{1+e^{\tau_k}}\delta_{jk}\right) = (1-\delta_{jk}) + \frac{e^{\tau_k}}{1+e^{\tau_k}}\delta_{jk} \tag{4}$$

where τ_k is the parameter that is estimated directly. We further define for δ_{kj} at $\delta_{kj} = 1$ such that

$$\rho_k = \frac{e^{\tau_k}}{1+e^{\tau_k}} \tag{5}$$

In the model that follows, the first attribute will be $-price$ so that in common with much of the literature, we specify a coefficient that is bounded in the positive domain. The parameter ρ_{kj} then becomes a multiplicative parameter for each of the marginal utilities in the following way:

$$\begin{aligned} \beta_{1j} &= \exp(\alpha_{1j})\,\rho_{1j} \\ \beta_{kj} &= \rho_{kj}\alpha_{kj}\ for\ k = 2, 3, ..., K \end{aligned} \tag{6}$$

where α is normally distributed. The inclusion of this element within our model means that we are employing a mixture of normals which means that our resulting posterior densities will not tend towards being "normal" in shape which would have occurred if we had simply employed the standard HBL. As such this specification has features in common with the classical flexible mixing distribution models introduced by Train (2016) and employed by Bazzani et al. (2017) and Caputo et al. (2018) as well as related Bayesian infinite mixture specifications used by Balcombe et al. (2017) and Ukpong et al. (2019). Further model details and the form of the hierarchical distributions used are provided in Appendix A.

Next, as is becoming common in the DCE literature, we estimate our model in WTP space. By specifying our model in WTP space, we obtain the well know and frequently cited benefits of estimating a model in WTP space (see, Train and Weeks, 2005). In addition,when employing a Bayesian approach to model estimation, it means we can employ prior information on the parameters that is more meaningful than if we estimated the model in preference space. Also, in setting our priors in this way we do not need to take account of scale within the priors and this will always vary between model specifications and data sets.

To specify the model in WTP space we achieve this by means of a simple transformation using the *price* coefficient. Thus, starting with the standard utility specification,

$$U_{ijs} = -\beta_{1j}x_{1js} + \beta_{2j}x_{2js} + ... + \beta_{kj}x_{kjs} + e_{ijs} \tag{7}$$

such that x_{1js} is the price attribute, we transform equation (7) as:

[3] ET data has also been used to attribute non-attendance in a number of papers such as Van Loo et al. (2018b).

$$U_{ijs} = \beta_{1j} \left(-x_{1js} + \frac{\beta_{2j}}{\beta_{1j}} x_{2js} + ... + \frac{\beta_{kj}}{\beta_{1j}} x_{kjs} \right) + e_{ijs} \tag{8}$$

With this transformation, we directly recover the marginal rates of substitution (MRS) and WTP estimates.

Finally, given that we are principally examining the relationship between our ET data and preference stability, we do not consider it necessary to undertake an extensive model selection exercise. Although model comparison exercises can contribute a great deal to a paper they are beyond the scope of the research being presented.[4]

5. Results

Our results begin by examining the full set of responses (i.e., n =99) in Sections 5.1 and 5.2 to examine if behaviour is consistent with that reported from previous ET studies. We then examine in Sections 5.3 and 5.4 only those responses for the repeated choice cards to consider preference stability and task complexity.

5.1. Descriptive Analysis of ET Data

We first examine proportional dwell defined using length of dwell and visual attendance. In the case of dwell time, we consider the relative proportion of dwell time for each attribute over the 24 cards completed by each respondent. As some of the attributes examined (i.e., Organic (ORG) and Quality Assurance (QAS)) in the DCE are either "absent" or "present" this means that there are blank spaces on the choice cards. It was apparent in these cases that survey respondents did not dwell on blank regions (it seems that respondents do not need to look to know information not there, i.e., they do not need to look at a blank region to know its blank). Therefore, these measures were modified so that if only one non-blank region appeared, the dwell was doubled for this attribute. Key summary are reported in Table 1.

Table 1: ET Summary Measures

	Proportional Dwell			Visual Attendance		
Attribute	Mean	Median	St Dev	Mean	Median	St Dev
Price	0.22	0.22	0.05	0.93	1.00	0.11
CoO	0.14	0.15	0.06	0.67	0.75	0.27
QAS	0.09	0.09	0.04	0.85	0.87	0.15
PQU	0.12	0.13	0.05	0.67	0.75	0.26
ORG	0.06	0.05	0.03	0.77	0.79	0.13
CAL	0.22	0.20	0.08	0.84	0.91	0.19
CoP	0.11	0.11	0.04	0.58	0.62	0.26

[4] All models are estimated using Stan (http://mc-stan.org/). This code should work irrespective of the platform within which Stan runs (R, Python, etc.). Although this data set required quite a high number of iterations to satisfy convergence, it was still many times faster than that required by other code such as Train (2009) Gauss routines that are commonly employed. The benefits of using Stan are several: first, it is compiled in C which is very fast; second, it uses Hamiltonian Monte-Carlo (HMC) which can be much quicker that Gibbs Sampling and/or Metropolis algorithms; and, third it is able to run multiple chains simultaneously, thus using multiple cores.

From Table 1, we first note that even correcting for blank regions on our choice cards, the attributes in questions still have the lowest levels of proportional dwell. Second, Price is highly visually attended and has the joint highest proportional dwell (i.e., it has high attention relative to other attributes). This is notable, as Price non-attendance is particularly vexing in the context of DCEs since the derivation of valid WTPs require Price attendance. While a small proportion of respondents (around 5%) only attended Price in around 50-60% of cases, the fact that average attendance was 93% and that median respondent attended Price in all 24 cards is an encouraging finding in our view, particularly in the light that in studies of SANA, reported rates have been much higher. Third, it is notable that Calories (CAL) also has high proportional dwell, nearly to the same degree as Price. However, Calories are not as visually attended as much as Price. We can also observe in the data that a few respondents spent almost 60% of their visual attention on CAL. Fourth, for some of the attributes such as QAS and ORG although they achieve reasonably high levels of visual attendance, they have quite small measures of proportional dwell. This indicates that certain bits of information are sufficiently salient that respondents do not need to dwell for long. And finally, it is worth noting that with respect to Country of Origin (CoO), Country of Production (CoP) and Product Quality (PQU) there are a small number of respondents that did not attend these attributes throughout the DCE.[5]

5.2. WTP Estimates and the Relationship with Visual Measures

Next we consider WTP estimates and we begin by examining the plots of distributions of the WTP estimates which are shown in Figure 3.

[Approximate Position of Figure 3]

What is evident from Figure 3 is the skewed, bimodal or multimodal nature of the estimates. These results have occurred partly as a result of the mixtures of normals that we have employed within our model specification. The mixture stems from the fact that we have conditioned our model parameters on the attribute attendance parameters that we have estimated. The results shown in Figure 3, are particularly stark in relation to the QAS estimates, that have a mass of respondents (around 20 to 25%) sitting on or around zero WTP, with another mode sitting above £1 for both Freedom Food and International Quality attribute levels. This is consistent with the notion that many respondents are primarily focused on a subset of the attributes, but not to the extent of totally ignoring all other attributes.[6]

Next, we consider the relationship between dwell duration and WTP. Based on the results reported in Table 1, we know that PQU was not dwelt upon to a great extent. However, the PQU attribute level, Premium vs Basic has the largest WTP of any attribute level, with Choice vs Basic also very large. Thus, absolute dwell duration does not reflect the importance of this attribute to respondents. We believe that this indicates that respondents were able to extract PQU information very quickly which would explain the relatively low dwell duration.

[5] Additional analysis on the extent to which different measures from the ET data are related and how they relate to attribute attendance are provided in Appendix B.

[6] Mean and medium attribute WTP estimates are provided in Table 2C in Appendix C.

We can examine this relationship further by fitting simple models to data on attribute estimates of WTP and proportional dwell measures. Thus, in Figure 4, we take the WTP by attribute and regress this against the logged absolute difference between least liked level of the attribute and the most liked level of the attribute for each individual.

[**Approximate Position of Figure 4**]

For each attribute shown in Figure 4, we provide the a p-value for null hypothesis of no correlation. We can see from this figure, a positive relationship between proportional dwell on the horizontal axis and our WTP measure on the vertical axis signifies that a higher level of utility is associated with an attribute if it is dwelt upon proportionally longer. The two lines in each of the panels of Figure 4 are quadratic lines fitted using the OLS estimator and a non-parametric (LOWESS) estimator. The significance values reported at the top of each panel are for the test of no-regression using the quadratic OLS specification. These models are simple, but sufficient to illustrate that we generally have, like Balcombe et al. (2017), a statistically weak but positive association between the attribute WTPs and proportional dwell. A person who looks at one of the attributes longer than another is more likely to care about the attribute to a greater degree, but the fact that people in general dwell on a given attribute for longer, does not mean that an attribute provides more utility in general.

5.3. Preference Stability and Dwell Duration

As explained for 50 of our participants, we given a set of repeated choice cards. That is, they received 12 cards sequentially, then received the same 12 cards again. Participants were not informed that they would receive repeated choices, and we believe that having faced 12 cards, it would be unlikely that respondents would notice that they were repeating the same choices. A full set of graphical results for length of dwell and visual attendance by reversal counts plus the relationship between choice reversals and WTP are presented in Appendix D.

We begin by examining the number of choice reversals our 50 respondents made. These results are shown in Figure 5.

[**Approximate Position of Figures 5**]

As we can see in Figure 5, 9 out of 50 respondents made no reversals (they were entirely consistent over the two sets of 12). The modal number of reversals was 2 (10 out of 12 were consistent) and 5 people made either 4 or 5 reversals.

Following Segovia and Palma (2020), we calculated the probability of making 0 to 12 random reversals. For somebody who had 5 reversals, the maximum number we report, the probability of making less than or equal to this number is approximately 38% falling to less than 2% for 2 random reversals. Second, we calculated Cohen's Kappa and we can report that 5 reversals, yields an estimate of 17% and for less reversals the percentage approaches zero quickly. Cohen's Kappa for each level of choice reversal is shown in Figure 5. Both sets of results indicate that the probability of making random choices in our data is relatively low.

Next, we examine the length of dwell by reversal count, where we have merged the 4 or 5 category of reversals given the small numbers making this many. What we find is that the average dwell on attributes such as Price, CoO, and CoP is significantly increasing with the number of reversals, and this is pretty much repeated across the board, although less significant for the other attributes. The results here are quite stark, and unambiguously show that those making reversals tend to dwell longer overall and on most attributes. This is an interesting and important finding as it implies that respondents are engaging with the choice tasks but it does not mean that they are able to make choices that imply that they have stable preferences.

Next, we can report that visual attendance by those who make more reversals is at least as high as those making less reversals. However, it is not generally statistically significant except for CoO and CoP at the 5% level of significance. In addition, CAL is the only attribute which shows a negative but statistically insignificant relationship between reversals and visual attendance. Overall these results show that respondents who made reversals generally dwell longer on attributes and visually attend the attributes at least as well as respondents with fewer reversals. We note, that it has been reported in the literature (e.g., Meißner et al., 2016) that attributes with greater importance to respondents generate greater attention. However, we observe that salience of an attribute (as inferred from the length of dwell) can mean that less attention is required to process an attribute even if the utility associated with an attribute is higher. The implication of this is that the length of dwell and visual attendance are not perfect substitute measures even if the measures are reported as being frequently highly positively correlated.

Finally, the relationship between reversal behaviour and WTP estimates by individual has been examined. The essential observation to be made, is that there is no clear monotonic pattern of increasing or decreasing WTP for most of the attributes. The clear exception is the scale parameter which decreases (significantly) as the number of reversals increase. This illustrates that from a modelling perspective that the random error (in terms of the Gumbel distributed error within the utility model) increases relative to the size of the systematic utility component. Thus, there is very little evidence that those respondents who reversed their choices display substantially different behaviour in terms of their underlying preferences. However, they were far more "stochastic" in making choices.

5.4. Dwell Time and Choice Complexity

We now examine what these results mean for RUMs, and respondent behaviour. Under the assumption of random utility we should see preference stability and no reversal of choices. However, even the most ardent RUM supporter is unlikely to suggest that respondents would or should never reverse their choices. Thus, optimistically, we would argue that the results show a relatively small proportion of respondents making more than 3 reversals out of 12. Pessimistically, we would remark that only 9 out of 50 were able to achieve preference stability which is slightly less than 20%. In comparison Segovia and Palma (2020) report values lower than this based on the results reported in their Figure 5.

Given the existing literature, it is fair to assume that respondents making choice reversals would be those respondents that had less engagement with the DCE, thus more likely to violate RUM. However, from the analysis of our ET data and WTP estimates, this is clearly not the case. Overall, respondents that made more choice reversals engaged well (i.e., high dwell time and visual attendance), if not better, with the DCE relative to those with less choice reversals.

Fundamentally, their utility was not particularly different to other respondents except that unsurprisingly they appear to have more "noisy" or less stable preferences.

Our interpretation of these results would be that the RUM may be a reasonable approximation of utility, but respondents do not have perfect knowledge of their preferences or are unable to articulate them. Furthermore, in many real world settings an individual will have experienced being genuinely undecided about choices, yet they are still able to make choices if required. Preference instability is not generally the product of lack of consideration by the respondents, but more likely to be the result of having similar utility for the different options as previously noted by Swait and Adamowicz (2001) and Meißner et al. (2016).

We now examine this likelihood graphically and econometrically. First, a plot of the dwell time on attributes by task throughout the sequence of the DCE is presented in Figure 6.

[**Approximate Position of Figure 6**]

This figure illustrates that there appears to be a slight downward trend in total dwell on each task as the experiment progresses, but with considerable randomness in relation to each of the attributes. This finding is consistent with results previously reported in the literature by Meißner et al. (2016) and Orquin et al. (2018).

Next, we examine econometrically the extent that dwell time declined over the sequence of choice tasks using a mixed random effects model. Our econometric model considers the extent to which dwell time is related to card complexity by introducing a measure of choice probability entropy similar to that proposed by Swait and Adamowicz (2001) and employed in subsequent studies such as Balcombe and Fraser (2011).[7] The model estimated is as follows:

$$\ln y_{it} = \mu_i - \lambda_i \ln(t) - \gamma_i \varepsilon_{it} + u_{it} \tag{9}$$

$$\begin{pmatrix} \mu_i \\ \lambda_i \\ \gamma_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \lambda \\ \gamma \end{pmatrix}, \Omega \right) \text{ and } u_{it} \sim N\left(0, \sigma^2\right)$$

where the dependent variable is total dwell on the *tth* card by the *ith* person ($\ln y_{it}$) and $\ln(t)$ is the log of time such that λ can be interpreted as an elasticity of dwell with respect to the number of choice tasks. In addition, ε_{it} is the choice probability entropy associated with each choice card, that is derived as follows:

$$\varepsilon_{it} = -\ln(p_{it})\, p_{it} - \ln(1 - p_{it})\,(1 - p_{it}) \tag{10}$$

where p_{it} was the probability that the *ith* person would choose the first option in the choice card t, as calculated by the HBL. A positive coefficient for (γ) implies that as the entropy measure goes up (the probability of choosing either of the two labels presented on the choice card tends towards a uniform probability of 1/2) then dwell would go up, suggesting that respondents

[7] Within the DCE literature a related approach to examining complexity has been employed by Olsen et al. (2011) and Uggeldahl et al. (2016). This approach works by *ex post* estimation of the expected utility difference between the chosen and non-chosen alternatives, and then examining how this estimate relates to variation in the ET data.

spend greater time looking at cards where the utility of each of the labels is very similar. The results are presented in Table 2.

Table 2: Model Estimates of Choice Complexity

	Parameter Estimates	Standard Error	Estimated σ^2	Standard Error σ^2	% <0
μ	8.635***	0.063	1.347	0.324	0.000
λ	0.145***	0.014	0.063	0.018	0.000
γ	0.420***	0.092	2.128	0.721	10.101

Note:***, **, * indicate significance at 1%, 5% and 10% level.

From the results in Table 2, we can see that there was an average decline in dwell ($\lambda = 0.145$) through the sequence of choice cards, implying that either respondents were either learning as the DCE progressed (and thus spending less time solving the problems) or that they are subject to a fatigue effect. While statistically significant this effect is relatively moderate, with a doubling of the number of cards (from 12 to 24) leading to a 10% drop off in dwell time. This appears to be consistent with a linear trend downwards, albeit subject to much card by card variation, rather than a non-linear one which would (might) be expected given adaptive or learning behaviour. However, without explicitly testing for either effect, we cannot ascribe the decline in dwell time to be either as a result of learning or fatigue. Regardless of the reason, the percentage of individuals estimated to have a negative coefficient, see the last column of Table 2, indicates that almost all individuals are dwelling less as the DCE progresses.[8]

Interestingly, the decline in dwell time was not a smooth process as shown in Figure 6, with the length of dwell devoted to a given attribute being variable across cards that was not dependent on the attribute levels of that attribute alone. Also, the average rate of decline in dwell time differed across attributes, with the most 'stable' attribute being Price which showed only a very small average decline in dwell time. Overall, our results in Table 3 and the ET descriptive statistics do not indicate a rapid deterioration in attention by respondents over 24 cards which was significantly longer than used in most DCEs.

Finally, turning to choice card complexity as captured by the entropy measure the estimates for the entropy coefficient γ is positive ($\gamma = 0.42$), which is consistent with people dwelling for longer for choice situations that are more difficult to decide between because they have similar utility. We also note that the percentage of those individuals estimated to have to dwell longer on higher entropy cards is almost 90%. The finding respondents dwell longer on cards that appear to have high entropy is in keeping with those previously reported in the literature by Pfeiffer et al. (2014). It therefore follows that respondents who are finding it hard to make decisions tend to look for longer than those for which the choice is clear. Furthermore, when this occurs there is an increased likelihood that the resulting choice can change if the same choice is offered again which indicates some degree of preference instability.

[8] We also re-estimated our HBL model by conditioning the variance of the WTP space model on the logged quadratic of the cards sequence following Balcombe et al. (2015). Results indicate that the logged quadratic term on the card sequence indicates a decrease in the decision noise over the range of the DCE, with a mean coefficient of 0.13 for the linear term and 0.001 for the quadratic. These results imply approximately a 10% increase in the length of the DCE leads to a 1.3% decrease in the standard deviation of the noise or a doubling of the experiment length decreases the noise by 13%. These results are in keeping with those we already reported in the manuscript in Section 5.4 and presented in Table 2. Importantly, based on model comparision criteria (i.e., WAIC) and WTP estimates there is no meaningful difference in model results generated.

6. Conclusions and Discussion

In this paper, we investigated the relationship between ET data and preference stability in a DCE in which we have repeated a high number of choice tasks. We have found that dwell time was positively related to the rate of choice reversals confirming findings previously reported in the literature in relation to response times and increases in error variance (e.g., Campbell et al., 2018). In addition, we also find that moderately low levels of dwell time do not indicate individuals with low levels of engagement with the DCE. We can draw this conclusion as we have also examined visual attendance data for the DCE at the attribute level. Equally, evidence of preference instability need not necessarily be assumed to indicate low levels of survey participant engagement especially as the degree of preference instability is positively related to dwell time. Most importantly, there appeared to be no substantive difference between respondents exhibiting high numbers of choice reversals compared to the rest of the sample in terms of their underlying preferences. It is simply that they are more likely to demonstrate preference instability and this takes the form of them making more "stochastic" choices especially when choice situations are more complex. Thus, we observe that respondents spend longer looking at tasks that offer choices yielding similar utility which makes any task more complex. Choices between options with similar utility pose more difficult choices for the respondent, and that respondents experience greater uncertainty about their choices under these circumstances. Respondents that experience preference uncertainty are more likely to make choice reversals.

Our analysis and results have yielded some important differences with the antecedent literature. For example, Meißner et al. (2016) report that respondent attention during a set of choice tasks is directed toward alternatives yielding higher levels of utility and as well as attributes yielding higher utility. We do not find this to always be the case. As we have reported, for some attributes the proportion of dwell time is relatively short yet they have high WTP estimates e.g., Organic. This result indicates that attribute salience can influence respondent engagement as measured by dwell time with the choice task and the relative effort required to understand the value to attach to a specific attribute.

From a practical perspective, knowing that more "difficult" choices take longer to make and that the actual choice made may not be "correct" is nothing new. However, we would expect there to be significant variation in DCE studies in large part due to the specific goods and/or services being examined. Much of the research on choice consistency of DCE that employs ET has considered goods that are likely to be well understood by participants (e.g., pizza, vegetables, coffee maker, laptops, beach holiday). The resulting degree of preference stability is likely to be a function of the content familiarity and as such more thought needs to be given to this issue when designing and employing DCE to value unfamiliar goods and/or services. Furthermore, it is also likely that context familiarity will result in some attributes being more salient due to prior product or service engagement.

Turning to the issue of conditions required to be met under random utility, it is unlikely that many respondents behave in a manner which is fully consistent with random utility, especially to the extent of processing all of the information all of the time. However, it is our view that such a requirement is extreme, and the degree of visual processing observed by researchers suggests that random utility models may be a reasonable approximation of respondent behaviour. In this DCE, we have found relatively high dwell time for Price. This, in conjunction with the fact that most attributes are visually attended most of the time by most respondents, bodes relatively well for the internal validity of DCEs. We would note, however, that we have observed in a number of other studies much larger rates of attribute non-attendance with regard to the price attribute, which raises particular problems for the analysis of DCEs in terms of producing meaningful WTP estimates. We believe it would be useful to return to and redo studies with high levels of SANA and observe to what extent this stems from weaker visual attention and/or attendance of price. From an economist's perspective, attendance of price is critical, and we believe that one of the best uses of ET is to ensure that DCE designs achieve high levels of visual engagement (dwell and/or attendance) of the price attribute before being administered in the field.

One particular issue that our analysis of choice task complexity and associated preference stability raises is with regard to the experimental design of DCE. Pfeffier et al. (2014) observed that taking account of (context) complexity is potentially an important task when designing a choice task. It is also understood that more complex experimental designs are likely to increase the cognitive demands placed upon respondents. For example, Johnson et al. (2017) observe that *"In some cases, complex designs may not lead to statistical improvements, for example due to trade-offs between statistical efficiency and respondents' cognitive capacity (or response efficiency)."* (p. 337). This complex relationship stems from the key objective of experimental design which is to maximise the information revealed by respondents about their preferences when making choices. It is also understood that utility balance maximises the information revealed by a specific choice. Yet, as noted by Regier et al. (2014), *"Applied research concludes utility-balanced designs increase the cognitive complexity of designs."*(p. 41). In fact, in a situation in which there is almost perfect utility balance the choice facing respondents is such that all alternatives are almost equivalent in terms of the utility derived. Clearly, in this situation respondents will struggle (it becomes almost impossible) to identify a preferred option and as such the random component in an econometric model will increase in magnitude. Thus, our findings support the point made by Johnson et al. (2017) that there needs to be a consideration of the trade-off between statistical efficiency and response efficiency on the part of researchers when design choice tasks. Even though this is not a new problem it remains and important feature of DCE experimental design that warrants further attention.

Another interesting feature of experimental design and how this relates to the ET choice literature is the difference in design methods employed within the DCE literature and the conjoint literature. Within the conjoint literature there appears to be is a clear preference for employing standard fractional orthogonal designs with little commentary regarding the efficiency of the design. For example, Meißner et al. (2016) state that they generated a set of randomized choice tasks that are approximately orthogonal and balanced by frequency of features. They also report that the other data sets they examine are not generated using any form of efficient design. This also the case for Yegoryan et al. (2019) who re-use old conjoint data sets that have been generated using orthogonal designs and random designs. In contrast within the DCE literature it is now commonplace to see researchers using efficient designs. For example, Uggeldahl et al. (2016) used a Bayesian D-efficient fractional factorial design. Although there is no doubt that either approach to experimental design allows for an examination of complexity the influence of different design criteria on any analysis warrants further attention.

Turning to future research, there would appear to be important links to explore between the DCE literature and the literature considering rational inattention as well as the use of response times. In both cases, these literatures are concerned with the process of decision making and not just the decision outcome. With regard to rational inattention, there is growing theoretical literature extending the original work of Sims (2003), such as Woodford (2014) and Caplin (2016). There are also studies, such as Cheremukin et al. (2015), that model and estimate the costs associated with making a choice given the benefits that result. By considering this trade-off during the process of making a choice there is an opportunity to improve how we econometrically model choice data in a theoretically robust framework. Similarly, with regard to response times, understanding how this data can be used to better understand choices not only relates to choice task complexity but also how response times can be used to derive underlying preferences (e.g., Konovalov and Krajbich, 2017). Importantly, our results support those reported in the response time literature (see Spiliopoulos and Ortmann (2018) for an excellent review of this literature) with regard to discrete choices. Specifically, a quick response to a discrete choice can occur regardless of the difficulty of the choice, whereas a slower longer response is almost always associated with difficult choices. Importantly, it is the difficult choices, as we have found, that yield stochastic responses (in our case preference reversals) and it is for this reason that these responses give rise to slow errors. This realisation should help to inform how response time data is used in DCEs and stated preference research more generally. In addition, the review of the literature provided by Spiliopoulos and Ortmann (2018) suggests that when using data collected without the imposition of time constraints (they refer to this as endogenous response time) that in order to extract the full potential of that data researchers will require, *"the use of procedural (process-based), rather than substantive (outcome-based), models of behavior."* (p. 387). Interestingly, Spiliopoulos and Ortmann (2018) observe that economic studies have yet to employ process-based approaches instead relying on dual-system models of Kahneman (2011) which they question as an approach to aid understanding of how decision are made.

Another research question that warrants further examination is the difference in choice task format and associated learning effects versus fatigue. In the choice tasks we employ we report minimal effects either learning or fatigue as captured in the speed of responses as the choice tasks proceed. This might be as a result of having devised a choice card that mimics a food label which differs to many other DCE studies in the literature. By taking this approach, we presented our options in columns and the attributes in both columns and rows. This is in contrast to the the likes of Meißner et al. (2016) were respondents compare attributes for each product in separate rows and subsequently report much greater learning or fatigue effects as they observe a much greater decline in response times. We also only employed two alternatives in contrast to Meißner et al. (2016) and it has been noted by Meißner et al. (2020) that the number of alternatives in a task influences potential learning or fatigue effects. Thus, although a previously research question with regard to visual design of choice tasks there is clearly for more research to understand the relationship between choice task format and respondent engagement.

Finally, an issue that emerges from the research presented is the extent to which it is even necessary to employ ET and instead simply rely on response time data which can be considered as a possible substitute for dwell time. There have already been questions raised in the DCE literature regarding the need to use ET partly because of the difficulty in scaling up data collection efforts and the use to which most ET can be made. In particular, Uggeldahl et al. (2016) questions the use of ET data compared to response times that can be obtained by the use of much less invasive technologies. However, in an exhaustive review of the ET literature, Zuschke (2020) concludes that there is still scope to use ET technology especially in relation to identifying salient information. As we have already noted, salience is an important feature in allowing respondents to quickly assess key pieces of information (i.e., attributes) and future research on food label design, especially with regard to healthy food choice, would benefit from knowing which parts of a food label are salient and which are not.

References

Adamowicz, W.L., Boxall, P.C., and Moon, A. (2009). Complexity in choice experiments: choice of the status quo alternative and implications for welfare measurement, **Australian Journal of Agricultural and Resource Economics**, 53(4):503-519.

Arentze, T., Borges, A., Timmermans, H. and DelMistro, R. (2003). Transport Stated Choice Responses: Effects of Task Complexity, Presentation Format and Literacy. **Transportation Research Part E**: 229-244.

Balcombe, K.G. and Fraser, I.M. (2011). Choice Experiments and "Don't Know" Responses. **European Review of Agricultural Economics**, 38(2): 171-192.

Balcombe, K.G., Fraser, I.M. and McSorley, E. (2015) Visual attention and attribute attendance in multi-attribute choice experiments. **Journal of Applied Econometrics**, 30(3):447-467.

Balcombe, K.G., Bradley, D., Fraser, I.M. and Hussein, M. (2016). Consumer Preferences Regarding Country of Origin for Multiple Meat Products. **Food Policy**, 64: 49-62.

Balcombe, K.G., Fraser, I.M., McSorley, E. and Williams, L. (2017). Examining the relationship between visual attention and stated preferences: a discrete choice experiment using eye-tracking. **Journal of Economic Behavior & Organization**, 144:238-257.

Bazzani, C., Palma, M.A. and Nayga Jr. R.M. (2018), On the use of flexible mixing distributions in WTP space: an induced value choice experiment. **Australian Journal of Agricultural and Resource Economics**, 62(2):185-198.

Bech, M., Kjaer, T. and Lauridsen, J.(2011). Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. **Health Economics**, 20(3), 273-286.

Börger, T. (2016). Are fast responses more random? Testing the effect of response time on scale in an online choice experiment. **Environmental and Resource Economics**, 65(2), 389-413.

Boxall, P., Adamowicz, W.L. and Moon, A. (2009). Complexity in choice experiments: choice of the status quo alternative and implications for welfare measurement. **Australian Journal of Agricultural and Resource Economics**, 53: 503-519.

Brazell, J. D., Diener, C. G., Karniouchina, E., Moore, W. L., Séverin, V., Uldry, P.-F. (2006). The no-choice option and dual response choice designs. **Marketing Letters**, 17(4), 255-268.

Campbell, D., Boeri, M., Doherty, E. and Hutchinson, W.G. (2015). Learning, fatigue and preference formation in discrete choice experiments. **Journal of Economic Behavior & Organization,** 119: 345-363.

Campbell, D., Mørkbak, M.R. and Olsen, S.B. (2018). The Link Between Response Time and Preference, Variance and Processing Heterogeneity in Stated Choice Experiments. **Journal of Environmental Economics and Management**, 88:18-34.

Caplin, A. (2016). Measuring and Modeling Attention. **Annual Review of Economics,** 8:379-403.

Caputo, V., Scarpa, R., and. Nayga R. M. (2017). Cue versus independent food attributes: The Effect of Adding Attributes in Choice Experiments. **European Review of Agricultural Economics**, 44(2): 211-230.

Caputo, V., Scarpa, R., Nayga Jr, R.M. and Ortega, D.L. 2018). Are preferences for food quality attributes really normally distributed? An analysis using flexible mixing distributions. **Journal of Choice Modelling**, 28: 10-27.

Carlsson, F., Martinsson, P., 2008. How Much is Too Much? An Investigation of the Effect of the Number of Choice Sets, Context Dependence and the Choice of Bid Vectors in Choice Experiments. **Environmental and Resource Economics**, 40(2), 165–176.

Carlsson, F., Morkbak, M.R. and Olsen, S.B. (2012). The First Time is the Hardest: A Test of Ordering Effects in Choice Experiments. **Journal of Choice Modelling**, 5(2): 19-37.

Chen, X., Gao, Z. and McFadden, B.R., 2020. Reveallling Preference Reversal in Consumer Preference for Sustainable Food Products. **Food Quality and Preference**,79, p.103754.

Cheremukhin, A., Popova, A. and Tutino, A. (2015). A Theory of Discrete Choice with Information Costs. **Journal of Economic Behavior and Organization**, 113: 34-50.

Clithero,J.A. (2018). Response Times in Economics: Looking Through the Lens of Sequential Sampling Models. **Journal of Economic Psychology**, 69: 61-86.

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

Gao, Z., and Schroeder, T.C. (2009). Effects of label information on Consumer willingness to pay. **American Journal of Agricultural Economics,** 91(3):795-809.

Green, P.E. and Srinivasan, V. (1978). Conjoint Analysis in Consumer Research: Issues and Outlook, **Journal of Consumer Research**, 5(2): 103–123.

Grether, D.M. and Plott, C.R. (1979). Economic Theory of Choice and the Preference Reversal Phenomenon. **American Economic Review**, 69(4): 623-638.

Hensher, D. A. (2006)a. How Do Respondents Process Stated Choice Experiments? Attribute Consideration under Varying Information Load. **Journal of Applied Econometrics**, 21:861–78.

Hensher, D. A. (2006)b. Revealing Difference in Willingness to Pay due to the Dimensionality of Stated Choice Designs: An Initial Assessment. **Environmental and Resource Economics**, 34:7–44.

Hensher, D. A., Stopher, P. R., Louviere, J. J. (2001). An exploratory analysis of the effect of numbers of choice sets in designed choice experiments: an airline choice application. **Journal of Air Transport Management**, 7(6), 373-379.

Hess, S., Rose, J.M. and Polak, J. (2010). Non-trading, lexicographic and inconsistent behaviour in stated choice data. **Transportation Research Part D: Transport and Environment**,15(7): 405-417.

Hess, S., Daly, A. and Batley, R. (2018). Revisiting consistency with random utility maximisation: theory and implications for practical work. **Theory and Decision**, 84:181–204.

Johnson, F.R. and Mathews, K.E. (2001). Sources and effects of utility-theoretic inconsistency in stated-preference surveys. **American Journal of Agricultural Economics**, 83(5):1328-1333.

Johnston, R.J., Boyle, K.J., Adamowicz, W., Bennett, J., Brouwer, R., Cameron, T.A., Hanemann, W.M., Hanley, N., Ryan, M., Scarpa, R., Tourangeau, R. and Vossler, C.A. (2017). Contemporary Guidance for Stated Preference Studies. **Journal of the Association of Environmental and Resource Economists**, 4(2):319-405.

Kahneman, D. (2011). **Thinking, Fast and Slow**. New York, Penguin.

Konovalov, A. and Krajbich, I. (2019). Revealed Indifference: Using Response Times to Infer Preferences. **Judgment and Decision Making**, 14(4): 381-394.

Krajbich, I. and Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. **Proceedings of the National Academy of Sciences**, 108(33): 13852-13857.

Krucien, N., Ryan, M. and Hermens, F. (2017). Visual Attention in Multi-Attribute Choices: What Can Eye-Tracking Tell Us? **Journal of Economic Behavior and Organization**, 135:251-267.

Lichtenstein, S. and Slovic, P. (1971). Reversal of Preferences Between Bids and Choices in Gambling Decisions. **Journal of Experimental Psychology**, 89: 46-55.

Louviere, J. J. (2004). Random utility theory-based stated preference elicitation methods: applications in health economics with special reference to combining sources of preference data, Working Paper at Center for the Study of Choice No. 04-001.

Mattmann, M., Logar, I. and Brouwer, R. (2019). Choice Certainty, Consistency, and Monotonicity in Discrete Choice Experiments. **Journal of Environmental Economics and Policy**, 8(2): 109-127.

Meißner, M., Musalem, A., and Huber, J. (2016). Eye tracking reveals processes that enable conjoint choices to become increasingly efficient with practice. **Journal of Marketing Research**, 53(1), 1-17.

Meißner, M. Oppewal, H. and Huber, J. (2020). Surprising adaptivity to set size changes in multi-attribute repeated choice tasks. **Journal of Business Research**,111: 163-175.

Morkbak, M.R. and Olsen, S.B. (2014). A Within-Sample Investigation of Test-Retest Reliability in Choice Experiment Surveys with Real Economic Incentives. **Australian Journal of Agricultural and Resource Economics**, 59: 375-392.

Pfeiffer, J., Meißner, M., Brandstätter, E., Riedl, R., Decker, R. and Rothlauf, F. (2014). On the influence of context-based complexity on information search patterns: An individual perspective. **Journal of Neuroscience, Psychology, and Economics**, 7(2): 103–124.

O'Donnell, M. and Evers, E.R.K. (2019). Preference Reversals in Willingness to Pay and Choice. **Journal of Consumer Research**, 45(6): 1315–1330.

Olsen, S.B., Lundhede, T.H., Jacobsen, J.B. and Thorsen, B.J. (2011). Tough and Easy Choices: Testing the Influence of Utility Difference on Stated Certainty-in-Choice in Choice Experiments. **Environmental and Resource Economics**, 49: 491-510.

Orquin, J. L., Ashby, N. J. S. and Clarke, A. D. (2016). Areas of Interest as a Signal Detection Problem in Behavioral Eye-Tracking Research, **Journal of Behavioral Decision Making**, 29(2-3): 103-115

Orquin, J. L., Chrobot, N. and Grunert, K. G. (2018). Guiding decision makers' eye movements with (un)predictable object locations. **Journal of Behavioral Decision Making**, 31(3): 341–354.

Orquin, J. L., and Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. **Behavior Research Methods**, 50(4): 1645-1656.

Regier, D.A., Watson, V., Burnett, H. and Ungar, W.J. (2014). Task Complexity and Response Certainty in Discrete Choice Experiments: An Application to Drug Treatments for Juvenile Idiopathic Arthritis. **Journal of Behavioral and Experimental Economics**, 50: 40-49.

Rigby, D., Burton, M. and Pluske, J. (2016). Preference Stability and Choice Consistency in Discrete Choice Experiments. **Environmental and Resource Economics**, 65: 441-461.

Sælensminde, K. (2002). The impact of choice inconsistencies in stated choice studies. **Environmental and Resource Economics**, 23(4): 403-420.

Segovia, M. S., and Palma, M.A. (2020). Testing the consistency of preferences in discrete choice experiments: an eye tracking study. **European Review of Agricultural Economics** (Online Early View).

Sims C. (2003). Implications of rational inattention. **Journal of Monetary Economics,** 50: 665–690.

Spiliopoulos, L. and Ortmann, A. (2018). The BCD of Response Time Analysis in Experimental Economics. **Experimental Economics**, 21: 383-433.

Swait, J. and Adamowicz, W. (2001). The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching. **Journal of Consumer Research**, 29:135-148.

Train, K. E. (2009). **Discrete Choice Methods with Simulation** (Second Edition ed.). New York, United States: Cambridge University Press.

Train, K. (2016). Mixed logit with a flexible mixing distribution. **Journal of Choice Modelling**, 19: 40-53.

Train, K.E. and Weeks, M. (2005). Applications of simulation methods in environmental and resource economics. In: Alberini, A., Scarpa, R. (Eds.), **DiscreteChoice Models in Preference Space and Willingness-to-Pay Space**. Kluwer Academic Publishers, pp. 1–16 (chapter 1).

Uggeldahl, K., Jacobsen, C., Lundhede, T.H. and Olsen, S.B. (2016). Choice Certainty in Discrete Choice Experiments: Will Eye-Tracking Provide Useful Measures? **Journal of Choice Modelling**, 20: 35-48.

Ukpong, I.G., Balcombe, K.G., Fraser, I.M. and Areal, F.M. (2019). Preferences for Mitigation of the Negative Impacts of the Oil and Gas Industry in the Niger Delta Region of Nigeria. *Environmental and Resource Economics*, 74: 811–843.

Van Loo, E.J., Grebitus, C., Nayga Jr., R.M., Verbeke, W. and Roosen, J. (2018). On the Measurement of Consumer Preferences and Food Choice Behavior: The Relation Between Visual Attention and Choices. **Applied Economic Perspectives and Policy**, 40(4): 538-562.

Woodford, M. (2014). Stochastic Choice: An Optimizing Neuroeconomic Model. **American Economic Review: Papers & Proceedings**, 104(5): 495-500.

Yegoryan, N., Guhl, D. and Klapper, D. (2020). Inferring attribute non-attendance using eye tracking in choice-based conjoint analysis. **Journal of Business Research**,111: 290-304.

Zuschke, N. (2020). An analysis of process-tracing research on consumer decision-making. **Journal of Business Research**,111: 305-320.

Figure 1: Example Choice Card

(without flag)

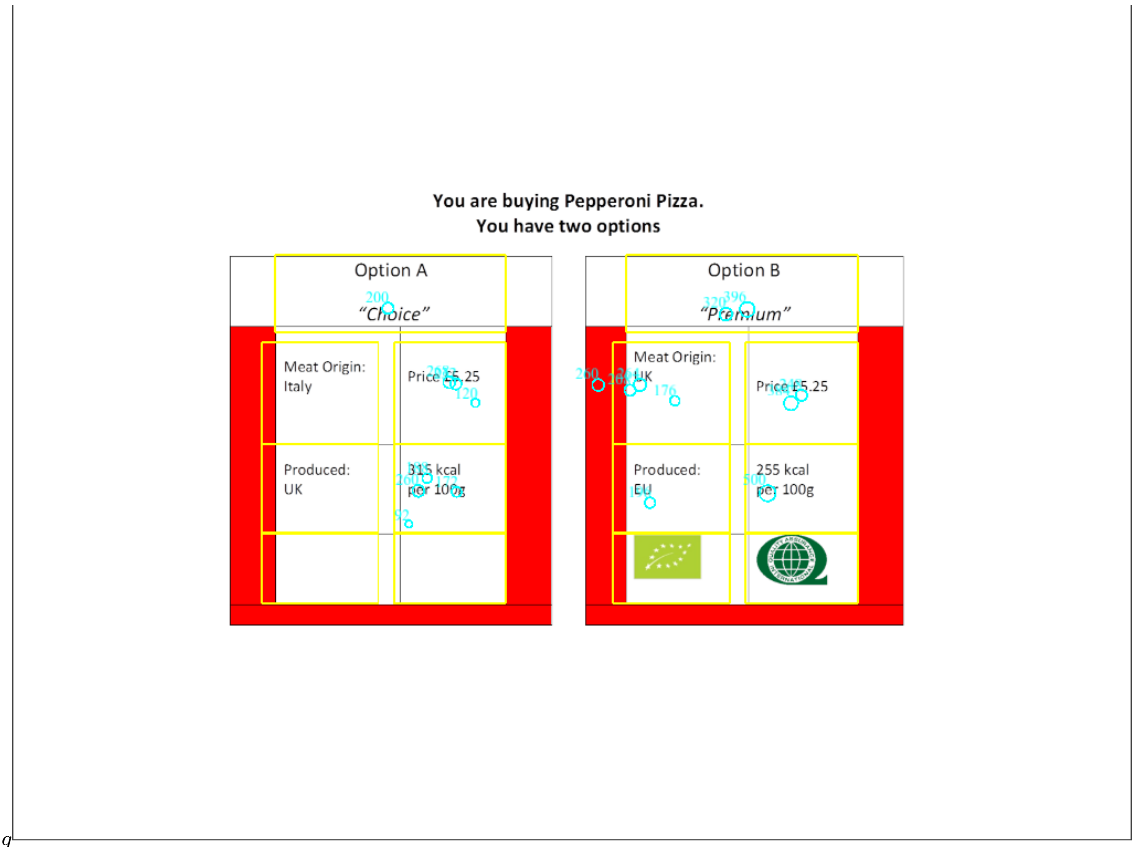


Figure 2: An Example Choice Card with Areas of Interest (AOI) Highlighted

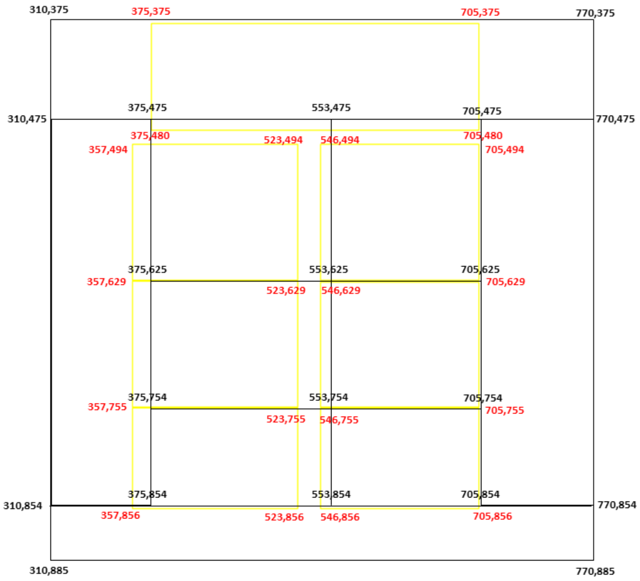


Figure 3: Attribute WTP Distributions

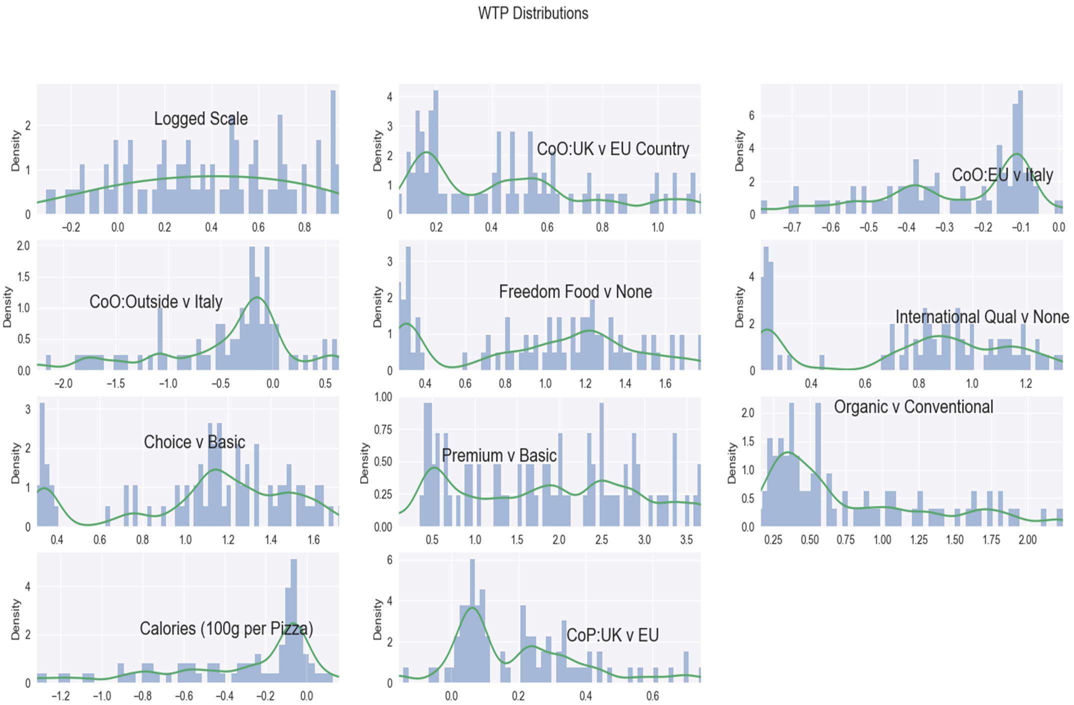


Figure 4: Relationship WTP and Proportional Dwell

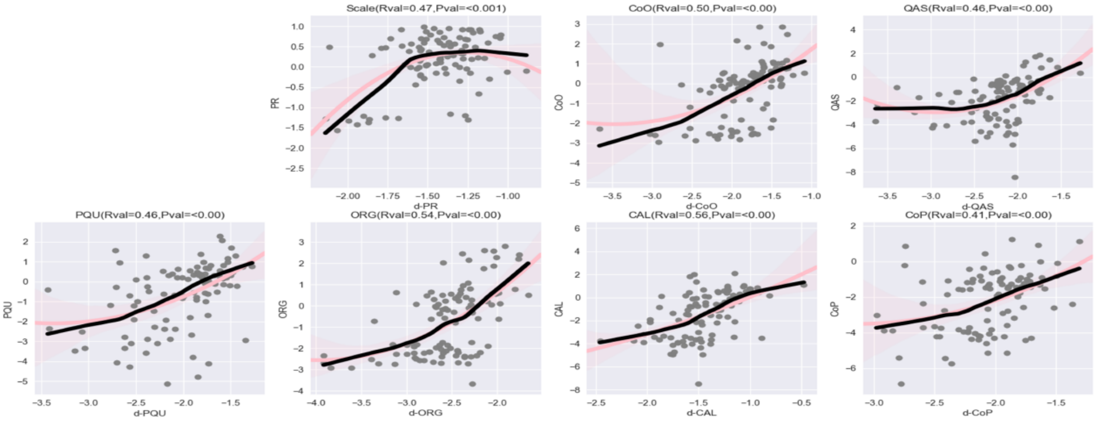


Figure 5: Number of Choice Reversals by Respondents

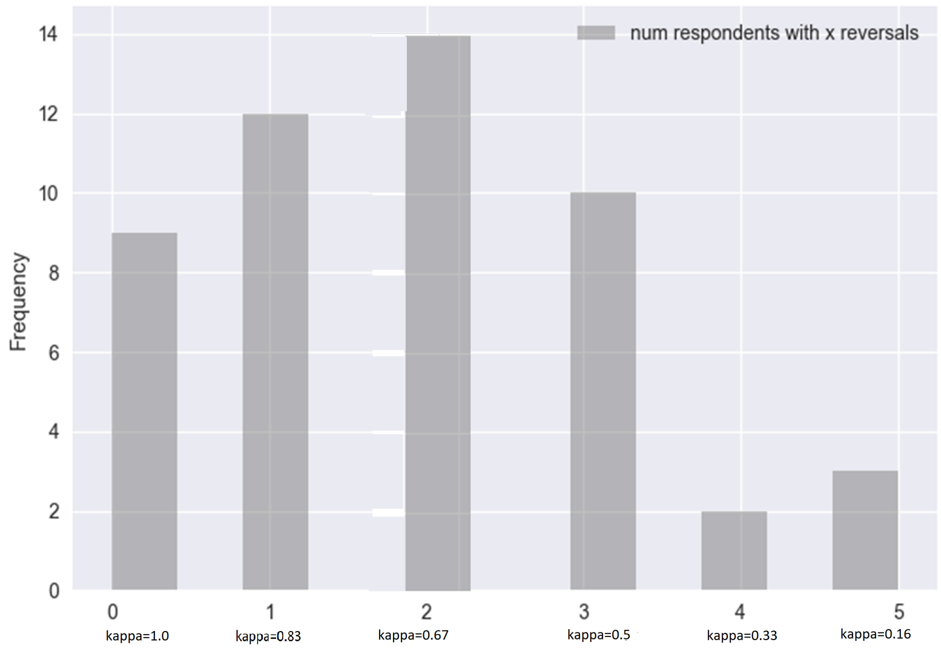
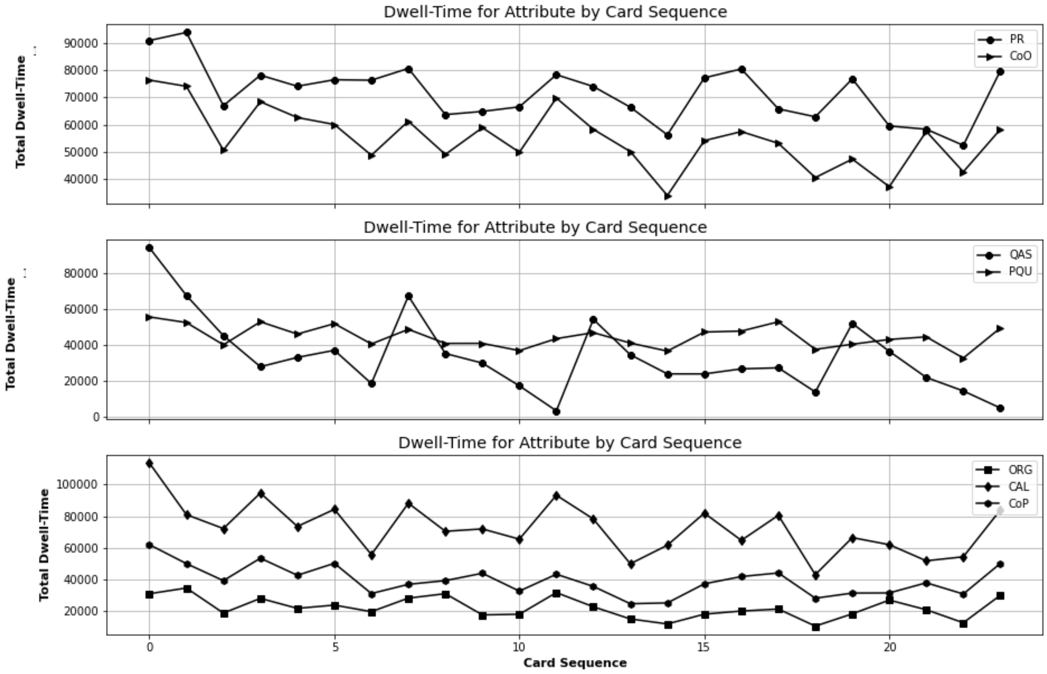


Figure 6: Dwell-Time by Attribute by Card Sequence



Online Appendices

Appendix A: Additional Econometric Details

We note that the existing approach employed in the literature with a shrinkage parameter is to assume $\rho_k = \rho$ for all k. However, allowing ρ_k to vary across attributes is possible, but estimating such parameters in a unrestrictive (non-hierarchical) way is not generally possible, because for many models respondents may have little (or even no) SANA for some of the attributes. In such circumstances estimating ρ_k will be non-identified (in the sense that the likelihood function will be invariant to its value). But, by estimating ρ_k using a hierarchical structure, the distribution for ρ_k for attributes where there is little or no SANA poses no problem since the posterior distribution of ρ_k for these parameters will be defined by the hierarchical distribution on which it depends. Importantly, Monte Carlo studies confirmed that if we have values for ρ_k that lead to alternative extremes in behaviour within one data set (i.e. $\rho_k = 1$ and $\rho_k = 0$ for $k \neq k^*$) then such behaviours will lead to posterior distributions for ρ_k, and ρ_{k^*} that are at the respective edges of the unit interval.

The full hierarchical structure is follows:

$$\begin{aligned} \alpha_{kj} &\sim N\left(\mu_k, \theta_k^{-1}\right) \text{ for } k = 1,2,3,...,K \\ \tau_k &\sim N\left(\tau, \eta^{-1}\right) \end{aligned} \tag{11}$$

along with:

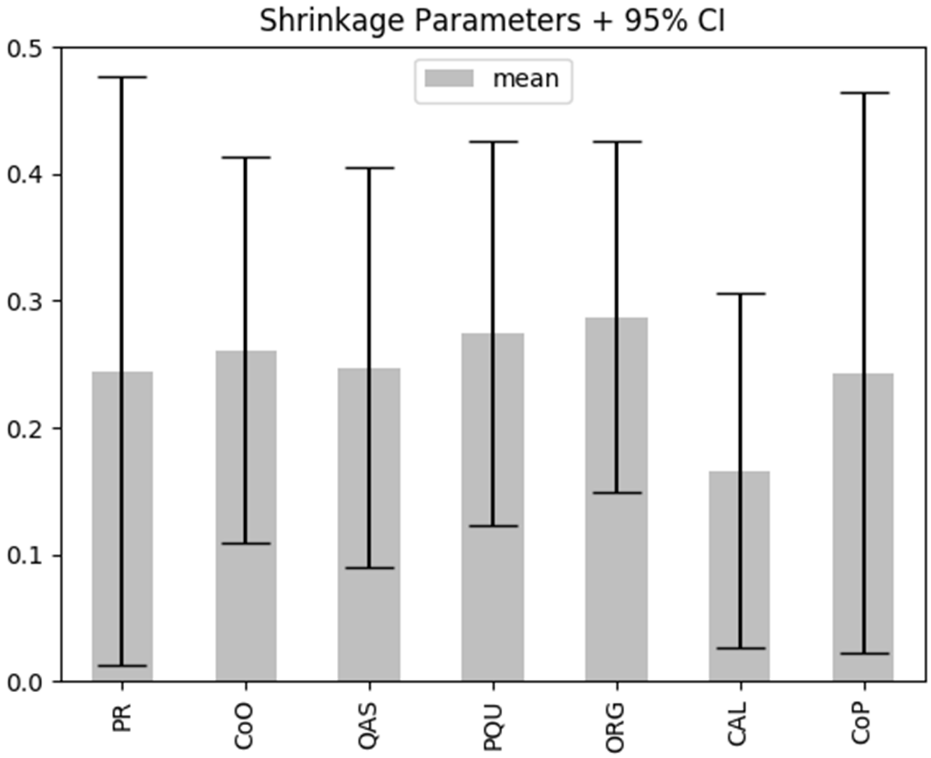
$$\begin{aligned} \mu_k &\sim N\left(\bar{\mu}_k, \bar{\sigma}_k^2\right) \text{ for } k = 1,2,...,K \\ \theta_k^{-1} &\sim G\left(\bar{a}_\theta, \bar{b}_\theta\right) \\ \tau &\sim N\left(\bar{\tau}, \bar{\theta}^{-1}\right) \\ \eta^{-1} &\sim G\left(\bar{a}_\eta, \bar{b}_\eta\right) \end{aligned}$$

where a bar above the parameter (e.g., $\bar{\mu}_0$) denotes a value that is set by the user.

The exact priors used in the empirical model presented are $\left(\bar{\mu}_0, \bar{\sigma}_0^2\right) = (0,1)$, $\left(\bar{\mu}_k, \bar{\sigma}_k^2\right) = (0,9)$, $\left(\bar{a}_\theta, \bar{b}_\theta\right) = \left(\bar{a}_\eta, \bar{b}_\eta\right) = (1,1)$ $\left(\bar{\tau}, \bar{\theta}^{-1}\right) = (0,25)$. We experimented with moderate changes in these priors, and obtained relatively small changes in the underlying WTPs and parameter estimates that we report.

Given our econometric specification, we examined the extent to which the estimates for our DCE attributes need to be changed as a result of SANA. Consider the attribute specific shrinkage measures (i.e., $\rho_k = \frac{e^{\tau_k}}{1+e^{\tau_k}}$), presented in Figure 1A. We can see that for those respondents stating that they did not attend a given attribute, that on average model estimates of stated non-attenders are around 20% of stated attenders, and on the basis of these findings, they do not seem to vary significantly across attributes. At least as far as this data set is concerned, employing a fixed shrinkage parameter as has previously been done in the literature (e.g., Balcombe et al., 2015) would not particularly bias results. However, given the sample size and number of respondents who state attribute non-attendance, the power to discriminate across attributes in this respect will be limited.

Figure A1: Attribute Specific Shrinkage Parameters



Note: These are 95% credible intervals as derived from the Bayesian HBL estimation

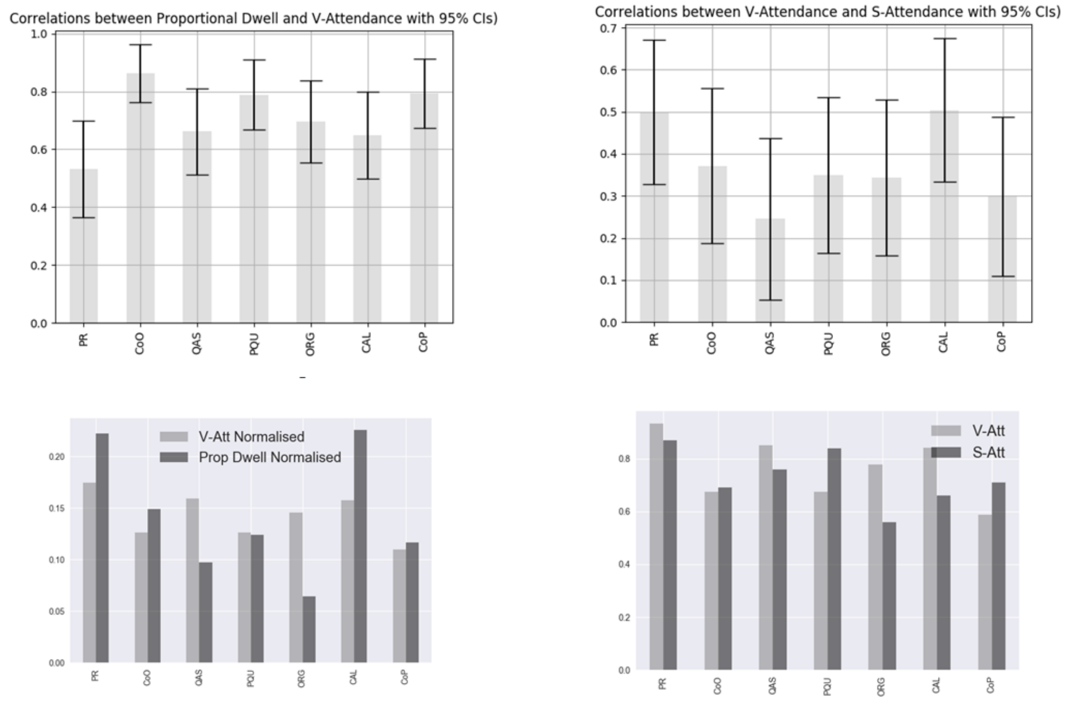
Appendix B - Alternative ET Measures

We have defined visual attendance to have occurred when an attribute is fixed upon (number of fixations) for all choices on a choice card. The relationship between dwell time and visual attendance is typically assumed to be such that either measure can be used in analysis. Here, we briefly examine the relationship between these measures for our ET data.

First, the two measures are significantly but moderately correlated across respondents. This is shown in Figure B1 below in panels B1.1 and B1.2. For example, we found that the higher the proportional dwell of an individual towards CoO, the higher their average visual attendance across all 24 choices, with a very high correlation close to 0.9. However, for CAL this measure is considerably smaller (i.e., correlation of around 0.65). Thus, for the case of CoO, many individuals who looked quite a while longer (proportionally) at this option, nonetheless did not visually attend this attribute to the same degree. Notably also, there is a large divergence for the two lowest attributes in terms of dwell time (QAS and ORG) which was not reflected in visual attendance. We conclude that this means that these attributes are salient in that respondents do not have to dwell for long to process the attributes even if they look at them (i.e., visually attend) many times.

Figures B1.3 and B1.4 compare the stated attribute attendance and visual attendance. As found by Balcombe et al. (2015, 2017) these measures are correlated but also quite distinct. Overall, these correlations, while all significantly positive, are moderate to small, with at most 25% of the variation in one of the measures (stated or visual attendance) being predictable from the other. In short, as expected, if somebody claims to have ignored an attribute, they are likely to have visually attended that attribute relative to somebody who claims not to have ignored it. However, the relationship is much weaker than one might expect, and the majority of survey participants who claim to have ignored an attribute have often fixed on all of the relevant information about that attribute in many or most of the tasks.

Figure B1: Measures of Attendance and Attention (With 95% Confidence Intervals)



Appendix C - WTP Estimates

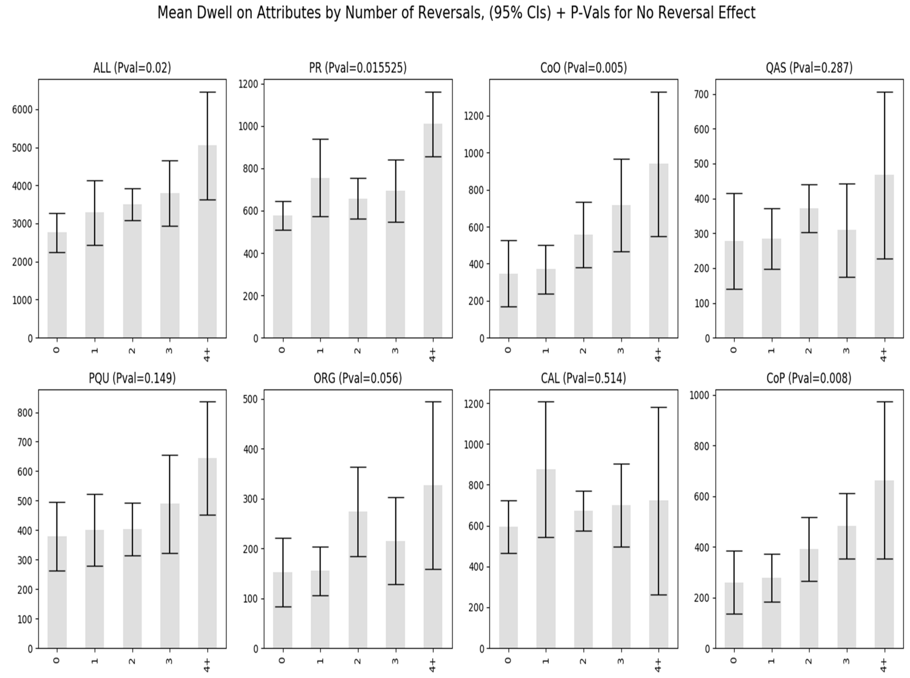
Table 2C: Attribute WTP Estimates

	WTP Estimates			
Attributes	Mean	St Dev*	25%	75%
CoO: UK vs EU	0.77	1.07	0.20	0.84
CoO: EU vs Italy	-0.42	0.52	-0.48	-0.12
CoO: USA vs Italy	-0.98	2.02	-1.09	-0.08
Freedom Food vs None	1.68	2.19	0.70	1.48
International Quality vs None	1.26	1.39	0.67	1.18
Choice vs Basic	1.67	1.73	0.99	1.53
Premium vs Basic	2.73	3.00	0.95	2.99
Organic vs Conventional	1.88	3.22	0.37	1.64
Calories (100g per Pizza)	-0.45	0.72	-0.60	-0.06
CoP: UK vs EU	0.34	0.60	0.06	0.36

Note:* St Dev - (Classical) standard deviations for the means for individuals derived from HBL model.

Appendix D: Additional Graphical Analysis of ET Data and Choice Reversals

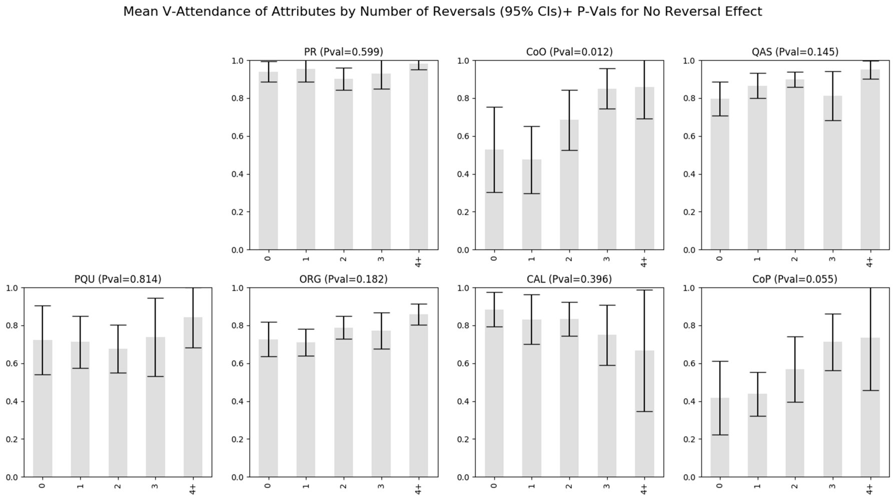
Figure D1: Mean Dwell by Attributes and Number of Choice Reversals



Note: Classical confidence intervals and classical P-values for a test as to whether the number of reversals is associated with the mean dwell.

This is a test for uniformity of the mean dwells over number of reversals.

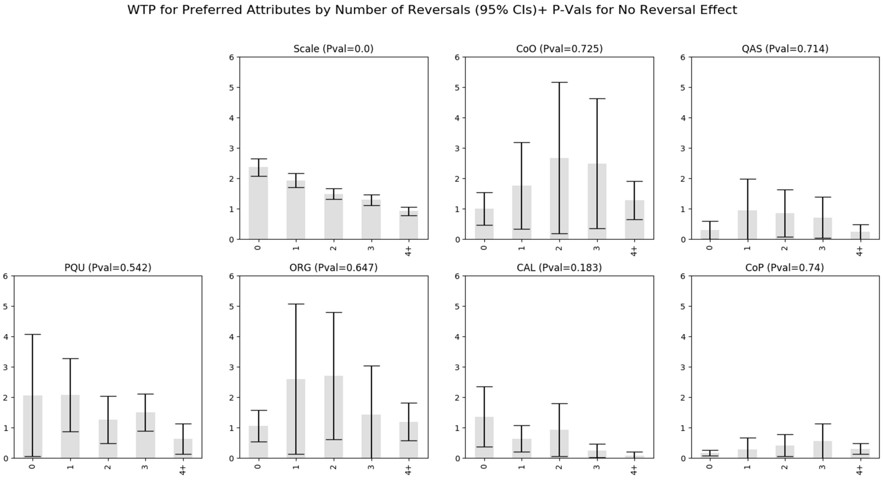
Figure D2: Visual Attendance by Attribute by Number of Choice Reversals



Note: Classical confidence intervals and classical P-values for a test as to whether the number of reversals is associated with the mean dwell.

This is a test for uniformity of the mean dwells over number of reversals.

Figure D3: Relationship Between WTP by Attribute and Number of Choice Reversals



Note: Classical confidence intervals and classical P-values for a test as to whether the number of reversals is associated with mean WTP for individuals (+ Scale).

This is a test for uniformity of the mean WTP over number of reversals.

Appendix E: Survey Instrument

See attached document for details