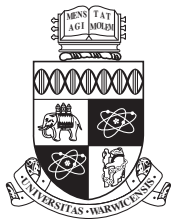


Beyond Introspection:

An exploration of the correspondence between directly and indirectly elicited preferences

by

Avri Avraham Moshe Bilovich

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Behavioural Science Group

Warwick Business School

January 2018

THE UNIVERSITY OF
WARWICK

Author's note:

This work was supported by the Economic and Social Research Grant ES/J500203/1 in partnership with Decision Technology Ltd.

Table of Contents

Acknowledgements

I would first like to thank my supervisor, Nick Chater for his continued help, encouragements and optimism which helped carry me through these interesting and sometimes challenging years. Our discussions always motivated me, and gave me something new to think about. I would also like to thank Daniel Read for stepping in when required and helping me stay on track. I must also thank Decision Technology, and in particular Keith Simpson, Rich Lewis and Henry Stott. Thank you for sharing your expertise with me and always being available. I am especially grateful to Keith whose inbox I periodically inundated with questions and who was always quick to answer. Sorry about the extra work I caused you.

I would also like to thank my friends and colleagues from Warwick Business School who have helped me to crystallise my ideas and have been used as a sounding board for new ideas, sometimes without even noticing. Most notably, I would like to thank Umar Taj and Marsha Kirichek for being there, especially during the last third of this thesis. I must also thank Edward Gardiner for distracting me from my core work with engaging, ethically aligned, applied projects.

My international friends have also been instrumental in my continued motivation for the last four years. Their confidence in me, coupled with their sometimes untimely distractions, have ensured that I kept enjoying myself while working. A special mention to the Biot Massiv': Maayan, Robin, Remy, Guillaume, Clem, Anna (et la famille), Yannick and Steph, as well as Amy and Leon. Having this many interested ears to explain my theories to has definitely helped me shape my ideas and express them more clearly.

There is no doubt in my mind that Rebecca Bendayan has had an important (and positive) effect on both my personal happiness and the success of this thesis. Thank you for supporting me and being a 'nudnikit' about all the things that I must improve in my work. You are wonderful.

Finally, I would like to thank my family for being always there: loving, supporting, motivating, distracting, … Knowing what to be at the right time. You have educated, motivated and pushed me forwards long before any of this PhD madness, and I have probably not told you enough how grateful I am for everything. I am also overjoyed to have been able to work while being with you during two 'once in a lifetime'. Thank you, mum, dad, Yarden, Yamit and Ehud, I love you all.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented (including data generated and data analysis) was carried out by the author except for the data collection in Chapter 5 which has been collected by Gallup Pakistan, but analysed by myself.

Abstract

This thesis aims to explore the limits of introspection, concentrating on the methodological aspect of preference elicitation. I use the comparison between direct and indirect elicitation methods, as well as their use in predicting choices, to explore when people are able to self-report their preferences. Three distinct contexts are used: the purchasing of habitual grocery products, the preferences of ethical consumers, and the political preferences of urban Pakistanis towards the USA.

The unreliability of direct elicitation methods, and the superiority of indirect methods in predicting choices that are uncovered in the first study are echoed in the context of ethical consumers. Indeed, even these more deliberative consumers, who reported an increased preference for ethical attributes, are shown to not be able to correctly report the drivers of their decisions. Conversely, the results of the third study indicated that urban Pakistani's preferences towards the USA were introspectively accessible: both types of elicitation methods were aligned in their results and were robust to short term attempts to change them.

This thesis contributes to the debate on the limits of introspection in several ways. Firstly, I demonstrate the inaccessibility of preferences for often experienced and purchased goods, even for more deliberative and attribute-conscious shoppers such as ethical consumers. I also demonstrate that unlike these preferences, political attitudes are more introspectively accessible. The work also presents a number of practical outcomes, such as the usability of lexicographic choice models with directly elicited preferences, as well as the plurality of ethical consumers' concerns. Finally, the robustness of political attitudes in Pakistan to actual pro-American advertisements indicate the need for a change of strategy in promoting the USA's positive impact in the country.

Abbreviations

AS: Anchored Scales (direct preference elicitation method)

CSS: Constant Sum Scales (direct preference elicitation method)

DCE: Discrete Choice Experiment

IAT: Implicit Association Test

LEX: Lexicographic choice model

WADD: Weighted Additive models

Chapter 1. Introduction

'*There are three things extremely hard: steel, a diamond, and to know one's self*' – Benjamin Franklin, 1750

'*If you don't know what introspection is you need to take a long, hard look at yourself*' - Ian Smith (comedian), 2015[1]

The difficulty of accessing one's preferences and internal states has been acknowledged for a long time. Psychological research has advanced the hypothesis that this is because people do not have introspective access to this information (Gopnik, 1993; Nisbett & Wilson, 1977a). Yet, every day we are still faced with numerous instances where introspection is required of us: from our partner asking us what we want for dinner, to filling up a customer satisfaction questionnaire. Are there not better ways to report our inner workings?

The understanding and identification of human preferences is central to many fields such as marketing, decision sciences and psychology. Applications are varied and include designing new products and predicting their popularity, predicting people's choices, and even assessing people's political attitudes. Although in practice, introspective methods, also known as direct elicitation methods, are still heavily used, indirect methods have risen in popularity (Wittink, Vriens, & Burhenne, 1994). The latter do not require people to have direct access to their preferences, and usually deduce those from choices or behaviour. However, the unequivocal superiority of these indirect methods has not been demonstrated (Green & Srinivasan, 1990; Sattler & Hensel-Börner, 2007).

This thesis aims to explore the limits of introspection, concentrating on the methodological concerns surrounding preference elicitation. In particular, the difference between direct and indirect preferences: when they diverge, when they converge and how can they be used to predict behaviour. This will be done through three studies, as presented in chapters three, four and five.

[1] As reported on the BBC: http://www.bbc.co.uk/news/uk-scotland-edinburgh-east-fife-34039927 (Last accessed: 24 November 2017)

Summary of the papers

In the following thesis, I present three studies which further explore how much introspective access people have to their attitudes and preferences, focusing on the differences between introspection based and non-introspection based preference elicitation methods. Each study has a distinct emphasis and is applied to a different domain, but all three are based on a similar methodology and further our understanding of introspection.

The first study (chapter three) investigated the hypothesis that preferences for familiar, often experienced, goods are introspectively accessible (Fischhoff et al., 1988). Hence, I focused on grocery products such as milk and eggs, to verify if people are able to reliably report their preferences, or if these are better uncovered through indirect elicitation. Methodologically, I used a number of direct, introspection based, preference elicitation and I compare them to an indirect one. This was conducted in three steps. First, I compared hypothetical choices in the experiment to ones taken in a supermarket setting, by using a UK-wide supermarket scanner dataset. I then compared the consistency of the attribute importance orderings elicited by the direct elicitation methods, and checked if these differed from the ordering of the indirect measure. The last step involved comparing the predictive validity of these preferences using two choice modelling frameworks: a lexicographic and a weighted additive. The addition of a lexicographical framework tests if these non-deliberative decisions can be better explained by non-weighted additive models.

The results confirmed the external validity of the study by showing a close correspondence between the hypothetical choices made experimentally and actual supermarket purchase behaviour (as measured by data from a large UK supermarket chain). Direct elicitation methods were not consistent, resulting in conflicting attribute importance orderings. These methods could not consistently identify even the most important attribute, for these familiar purchases. Finally, although the lexicographic framework coupled with a direct elicitation method showed promise in predicting choices, the unreliability of direct elicitation concedes a clear advantage to indirect elicitation and weighted additive models for identifying the drivers of people's behaviour and predicting their choices.

The second paper (chapter 4) contributes to the general aims of the thesis by exploring the degree to which direct and indirect preferences converge for ethical and less ethical consumers. This paper tests the hypothesis that ethical consumers have more introspective

access to their preferences, especially those which contain ethical attributes. Another aim of this study was to verify to what extent self-reported ethical consumers make choices that are consistent with their stated beliefs. This involved therefore comparing the differences between the choices of ethical and less ethical consumers. I therefore divided the previously collected responses into three groups, based on their self-reported frequency of ethical purchasing: ethical, occasionally ethical and non-ethical consumers. Although the method was similar to the one presented in the first study, the analysis method was very different. First, I described the demographical and attitudinal characteristics of each group, ensuring that these were consistent with their classification. Following which, the within and between group preference homogeneity was calculated. Lastly, a comparison between the direct and indirect preferences was conducted for all groups.

The results indicated that general attitudinal beliefs, as well as the demographics, correctly identified the three groups in terms of their ethical purchasing habits. The product-specific preference heterogeneity of each group revealed the plurality of ethical concerns: these were the least homogeneous in terms of their preferences. An attitude-behavioural gap was highlighted for all groups, by the comparison between the direct and indirect preferences. Unlike what was hypothesised, this was not driven by ethical concerns. The study highlighted the similarity between the choices of the three groups, thus suggesting the limited impact of general ethical concerns on specific choices in this context.

The final study (chapter 5) set out to generalise the previous findings to the domain of political attitudes. As familiar decisions were hypothesised to be more introspectively accessible, so did political attitudes. Therefore, this study aimed to verify if political attitudes, in this case pro or anti American ones in Pakistan, are introspectively accessible. Given the efforts of the USA in attempting to change the American sentiment in Pakistan, a secondary aim of this research is to verify if, in this context, stated preferences are more malleable than indirectly elicited ones. Using a similar methodology as in the other studies, I measured attitudes towards a number of countries, including the USA, using direct as well as indirect measures. Additionally, participants were randomly divided across three conditions, each one with a different advertisement which was showed after the first direct elicitation method. Overall, the results confirmed that attitudes towards the USA are negative in Pakistan, and showed that they were equally robust to short term attempts at attitude change through

advertisement. Additionally, direct and indirect attitudes were aligned, indicating that, unlike familiar decisions, such attitudes are introspectively accessible.

Contributions

These studies taken together contribute to the debate on the limits of introspection in several ways. Firstly, they demonstrate limited introspective access to preferences of habitual products. These often experienced products are expected to be more easily accessible (Fischhoff et al., 1988). The gap between attitudes and choices is also highlighted for these products, even for ethical consumers who should be more deliberative in their choices. Conversely, political attitudes are shown to be much more available to self-knowledge, and not liable to short term changes. In addition to furthering our understanding of introspection, this work has a number of useful practical outcomes. The empirical tests of lexicographic and weighted additive models, showed that although lexicographic models, which are easy to parameterise from direct elicitation methods, can be applicable in some contexts, indirect methods give more reliable prediction results. On the other hand, when uncovering political attitudes, the third study validates the introspective methods used by global surveying companies such as Gallup or the Pew Research Centre, by showing that they align with more complex indirect methods. The work in Pakistan has also provided a template for using more sophisticated indirect preference elicitation techniques in developing countries. The implementation of a discrete choice experiment using the open data kit can be generalised and translated to other domains. Finally, the comparison between ethical and non-ethical consumers confirmed the plurality of ethical concerns, and the limited impact of general ethical considerations on specific grocery choices.

The rest of the thesis is structured as follows. In Chapter 2, I review the evidence surrounding the limits of introspection, before presenting a general overview of preference elicitation methods and the choice prediction models which use them. This chapter concludes with the specific aims of the thesis, in light of the preceding discussion. Chapters 3, 4 and 5 respectively present each of the three studies, and can be read in isolation. Chapter 6 concludes with a general discussion and suggestions for further work.

Chapter 2. Literature review

This chapter provides an overview of the theoretical and practical background of the thesis. It is divided into three parts. The first one will expand on the notion that people have limited introspective access to their self-knowledge, following which a discussion of preference elicitation techniques will present the framework of direct (introspection based) and indirect methods for preference elicitation, as well as the choice prediction models which can be derived from them. The final section delves into the discussion about the nature of these measured preferences and the implication that this has on preference elicitation.

Introspection

The degree to which people are able to access their internal states has been long debated in psychology. Nisbett and Wilson (1977a) famously presented evidence against people's introspective access. The first examples showing this lack of introspective access can be found in misattribution studies. Misattribution studies use contextual cues to make people assume a link between two different processes. For example, Dutton and Aron (1974) asked an attractive female interviewer to approach people on a fear-arousing (less stable) and a non-fear-arousing bridge. The responders were asked to complete a Thematic Apperception Test indirectly to measure sexual arousal and were given the contact details of the interviewer if they wanted to gather more details about the survey. The latter was used as a behavioural measure of attraction, i.e. the number of people who would subsequently contact the researcher. The hypothesis of the study was that the arousal resulting from the fear-inducing bridge would be misattributed as sexual attraction towards the interviewer, and was corroborated by both measures. Another example of misattribution is given by Nisbett and Schachter's (1966) experiment in which two groups were exposed to a sequence of steadily increasing electric shocks, after being given a placebo pill. One group was told that the pill would produce heart palpitations, breathing irregularities, etc. (all common side effects of mild electric shocks). This group, which attributed the effects of the electric shocks on the pill, withstood significantly stronger electric shocks. Interestingly, participants expressed no awareness that they attributed these effects to the pill, instead giving other justifications as to why they resisted better to the electric shocks.

In perception, it is also well accepted that some things, even within our field of vision, are not attended to and slip outside of conscious retrieval. The now-famous "gorilla" experiment (Simons & Chabris, 1999), in which participants failed to notice a person dressed as a gorilla crossing their field of vision while they were attending a dynamic scene, is one such example. One of the reasons this occurs is because people are occupied with other cognitively challenging tasks, such as counting the number of basketball passes in this case. It could therefore be that the limited introspection which people exhibit is caused by similar distractors. Yet, there is evidence that careful deliberation and justification of preferences leads to lower quality decisions (Wilson & Schooler, 1991) and less choice satisfaction (Wilson et al., 1993). Deliberation can also cause attitude change, and increase the dissociation between attitudes and behaviours (Wilson, Dunn, Kraft, & Lisle, 1989). Hence, unlike visual perception, even when actively trying to interpret our internal states or to attend to the reasons behind our choices, we are unable to access them.

In general, people do give reasons for why they have made a specific decision. These reasons often do not contain all the drivers of the behaviour. In fact, there is a possibility that these reasons are constructed after a decision is made, as portrayed by the choice blindness phenomenon (Johansson, Hall, & Sikström, 2008; Johansson, Hall, Sikström, & Olsson, 2005). Building on people's perceptual inattention to change, Johansson et al. (2008) showed that it was possible to make people think that they made a different choice from the one they actually did. After making a binary preference choice between pictures of faces, participants were presented with the picture that they did not choose, as if they did make that choice, and asked to justify their choices. Although the picture was not one that they have chosen, people still gave reasons for why they have made that choice. The inattention to change persisted in a post test recall task where participants remembered the manipulated choice as their own. Thus, in some cases, people's access to their own decisions is limited, and their justification for their choices is constructed post-hoc.

The lack of introspective access to one's own mental processes and constructs, such as choices and preferences is also extended towards one's perception of oneself. For example, people tend to over-estimate their abilities (Kruger & Dunning, 1999) and their health risks (Dunning, Heath, & Suls, 2004) while underestimating how much their moderate beliefs shift (Wolfe & Williams, 2017). In addition to beliefs, people also seem to have limited access to the veracity of their intuitions (Leach & Weick, 2017). It has also been suggested that self-

perceptions and others' perceptions of oneself are often at odds, with insights coming from other people being more predictive of behaviour (Kolar, Funder, & Colvin, 1996; S Vazire & Carlson, 2011). Self-knowledge is therefore better attained by looking through how one is viewed by others rather than by introspecting (Wilson & Dunn, 2004). Carruthers (2011) refers to the system which is employed to introspect as the mindreading system, and posits that it is the same system that is used to reason about other people. He maintains that this system does not have much additional information when reasoning about oneself, and is therefore reliant on interpreting sensory evidence. This Interpretive Sensory-Access theory of self-knowledge takes into account the previously mentioned evidence for lack of introspective access and makes a bold claim: we deduce our attitudes and preferences using the same processes that are used to deduce other people's attitudes and preferences. The ease of misattribution of internal states is proof that the additional information that we possess about ourselves is inexact, noisy and interpretative by nature. This additional information that we have about ourselves can therefore sometimes be more of a hindrance rather than an advantage in deducing our internal states, which explains why other people's deductions about ourselves are more accurate (Kolar et al., 1996).

A final notion that is useful in understanding introspection is the illusion of exploratory depth (Rozenblit & Keil, 2002). This principle refers to how people feel like they understand complex phenomena, until they are asked to explain it in more detail. Rozenblit and Keil (2002) used mechanical phenomena, such as a zipper of a flush toilet, which participants were familiar with but failed to explain the working of, to identify this illusion. Similarly, people feel that they have more access to their decision making processes rather than all their mental events: this has been suggested as another example of the illusion of explanatory depth (Kozuch & Nichols, 2011). Hence the more we try to access our preferences or mental processes, the more we appear to be faced with the limits of our introspection.

Preference elicitation

Preference elicitation methods are used to understand what aspects, or attributes, of products drive people's choice. For example, when talking about a mobile phone, the attributes might include: size, brand and price. In turn, these attributes have different attribute levels. For brand, these are the specific brand names. When eliciting preferences, one must

identify not only the importance of the attributes, but also the importance of the levels within each attribute. For convenience, unless specified otherwise, the methods discussed in the following are applicable to both the importance of the attributes and their levels. Preference elicitation is usually divided into two categories: direct methods, which rely on introspection and indirect ones, which do not.

Direct preference elicitation

Direct preference elicitation, consists of directly asking people to report their preferences. This can be done in an unstructured way (e.g. Ding et al., 2011), using free text or open ended questions, which are often used in focus groups or interviews. Alternatively, a more structured approach commonly used in surveys requires people to rate or rank pre-identified attributes in their order of importance. The attributes in question are either deduced from the product category, or identified using unstructured methods. For example, the simple metric scale requires respondents to rate each attribute in turn, typically by giving it a numerical value between 0 and 10. With no further constraints on how people answer, responses tend to be grouped at the high end of the scale (Meyer, 1999), which is problematic as the informational content of the elicitation is lessened if many of the attributes are identical. In an attempt to provide more differentiation between the attribute ratings, the Magnitude Scaling technique (Lodge, 1981) provides respondents with an anchor, such as a neutral midpoint on the scale (5 in our previous example). Some approaches favour more than one anchor, such as Huber's (1974) client-explicated parameter estimates. This method uses two anchors, respectively the least and most important attributes, to anchor both the lower and higher end of the scale. Building on this approach, Srinivasan's (1988) self-explicated approach eliminates unacceptable attribute levels before following Huber's methodology. This enables the elicitation of only the relevant attributes or levels and thus can reduce the effort required by participants. Another way to ensure the differentiation between the attributes is by making it more difficult for respondents to give the exact same score to the attributes. The constant sum scales (CSS) method achieves this by requiring respondents to distribute a fixed number of points (usually 100) among the different attributes. Alternatively, respondents can be asked to simply rank the attributes, which also forces them to differentiate between the importance they assign each of them. Although very easy to respond to, this method makes the response scale ordinal. Therefore, there is no longer a way of capturing if a respondent wants to say that price is twice as important to them as size, for

example. Still, there are psychological theories which state that ordinal scales are enough to explain people's decision making process (e.g. Stewart, Chater, & Brown, 2006; Tversky, 1969), justifying the use of such elicitation methods.

Direct preference elicitation methods have been criticised for a number of reasons. Firstly, they rely on people's introspective access to their preferences, an issue which has been widely disputed in psychology as discussed previously. The dependency of responses on the scale provided has also lead some to criticise these methods. For example, Schwarz et al. (Schwarz, Hippler, Deutsch, & Strack, 1985) showed that using more high (up to 2 1/2h, 2 1/2h to 3h, …, more than 4 1/2h) rather than low (e.g. up to 1/2h, 1/2h to 1h, …, more than 2 1/2h) categories on a scale resulted in an increased television viewership reporting. This also affected the respondents' estimation of other people's behaviour in the same direction. Frequency scales have been shown to similarly affect behaviour reporting in other domains such as health (Schwarz & Scheuring, 1992) or sexual behaviours (Tourangeau & Smith, 1996). Another result that complicates the usage of direct elicitation methods is anchoring (Tversky & Kahneman, 1974), which occurs when respondents adjust their responses compared to an arbitrary value given beforehand. This is exemplified by a series of experiments by Ariely, Lowenstein and Prelec (2003). In one of these experiments respondents were asked to write the last two digits of their social security number, before valuing a product, such a cordless keyboard or some Belgian chocolates. Respondents who wrote a higher number, also indicated a higher willingness to pay for the products. Similar results were uncovered by subsequent experiments involving the pricing of hedonic (positive or negative) experiences, such as listening to unpleasant sounds after being exposed to an anchor. These anchoring results have also been replicated in the context of reporting one's own or estimating other people's behaviour (Cheek, Coe-odess, & Schwartz, 2015) and will be expanded upon in a subsequent section.

Hence, if people's answers to direct elicitation questions are affected by the response scale or the context, there is a strong possibility that they are not actually accessing their preferences, but constructing them when they need to (Lichtenstein & Slovic, 2006). Even if preferences are accessible and not constructed, other effects such as the social desirability bias (DeMaio, 1984) can limit direct elicitation answers. In this case, respondents may knowingly give a different answer to a question in order to portray themselves in a more positive light. More generally, Schwartz and Oyserman (2001) decompose the elicitation

process into 5 steps. Respondents need to understand the question, recall or access the answer, use some inference or estimation before mapping the answer onto the response format. The final step involves editing the answer for social desirability. Anchoring and response scale effects affect the third and fourth step, but there can be issues at each of these stages (see Schwarz & Oyserman, 2001). Despite these limitations, some of which can be resolved with careful design, direct elicitation methods are still widely used by practitioners. Their benefits include being easy to design and cheaper to run, especially in studies with large number of attributes (Sattler & Hensel-Börner, 2007). Indeed, when the number of attributes or levels is extremely large, direct elicitation methods, or methods which partially rely on direct elicitation may be the only ones feasible. Responses to direct elicitation methods are also often faster to collect and less straining on respondents than indirect ones.

Indirect Preference Elicitation

Indirect preference elicitation methods have evolved as a way to circumvent the critiques of these direct techniques. These methods do not ask respondents to value the importance of attributes. These are instead deduced from their choices or behaviours. The most well-known of these techniques is Conjoint Analysis (Green & Rao, 1971) and its variants (Green & Srinivasan, 1990). The original conjoint analysis task involved rating the relative attractiveness of a number of options. For instance, respondents would be faced with the description of 4 different breakfast cereals and would give each a rank order (between 1 and 4) dependent on their preferences for the products. This ordering is then fitted, for example by a mixed model (Næs, Brockhoff, & Tomic, 2010), in order to get a numerical representation of the importance of each of the attributes. A different approach to the rating based conjoint analysis is to look specifically at people's choices, rather than the full ordering. Hence, instead of people having to put all the cereals in order of preference, they are asked which one they would choose or buy. This approach, known as choice based conjoint or discrete choice experiment (DCE; Louviere, Hensher, & Swait, 2000), is considered more realistic, as it mirrors how consumers make purchasing decisions in the real world. On the other hand, if individual specific models of choice are required, then DCE might engender some decision fatigue as respondents will need to make repetitive choices from different choice sets. The number of choices to make depend on the design of the DCE including the number of options to choose from, the number of attributes and levels of each option, and if individual preferences are required or not (Hensher, Rose, & Greene, 2005).

Theoretically, DCE's are based on Luce's choice axiom (Luce, 1959) and random utility theory (McFadden, 1974). Luce's choice axiom, also referred to as the independence of irrelevant alternatives (IIA), posits that the probability of choosing a specific alternative over another should not be affected by additions or removals from the choice set. For example, the probability of choosing one milk carton over another, should not be affected by the inclusion of an additional carton. The utility, referred to by random utility theory, can be defined as the measure of satisfaction that a person receives, in this case from the good that is chosen, and is therefore used to identify preferences. At its core, this theory states that individuals make choices that maximise their utility. The utility associated with a specific good is comprised of a linear combination of the utilities of each attribute of said good. For example, when thinking about fresh milk, the type of the milk (semi-skimmed, skimmed, …), its price, its brand and its quantity all contribute to the total utility of a specific fresh milk carton. More formally, the utility U_{ij} that individual i associates with a specific option j within a choice set C is written as $U_{ij} = V_{ij} + \varepsilon_{ij}$, where $V_{ij} = f(X_j, \beta)$ is the systematic portion of the utility composed of a function of the vector of product attributes (X_j) and individual i's utility of these attributes (the vector of coefficients β), and ε_{ij} is the stochastic component. The stochastic component captures all unobservable attributes influencing choice and any measurement errors and is thus referred to as the random error component. Hence, an individual i will choose option n if $U_{in} > U_{ij}$ $\forall$ j $\in$ C, i.e. if out of all the options available in the choice set option n has the maximum utility. If we assume that ε_{ij} are independent and identically distributed, the probability that individual i would choose product j can be written as the standard multinomial logistic regression model:

$$Prob(choice = j) = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}}$$

The betas in the systematic portion of the utility function represent the preferences of the different attributes, relative to each other. For categorical attributes with multiple attribute levels, such as "type of milk" which can be skimmed, semi-skimmed or whole, these betas represent the utility of each individual level. Hence, in order to deduce the relative preference of the attribute as a whole, one must use the partial R-squared or the log-likelihood of the resulting model (Louviere & Islam, 2008). More specifically, the contribution of each attribute to the log-likelihood of the full model can be used as a proxy for its importance. The calculation is as follows: each attribute is removed one by one from the full model and the

difference between the log-likelihood of the full model and the omitted attribute model is recorded. This value, for each attribute in turn, is then divided by the sum of all the other ones. This results in a value representing the attribute's relative contribution to the log-likelihood of the full model, regardless of the number of attribute levels. Hence this method allows, for example, an easy comparison between the relative importance of price (as a continuous attribute), type of milk (a discrete attribute with more than 3 levels), and if the milk is organic or not (a discrete attribute with two levels). The multinomial logistic regression model is not the only discrete choice model that can be applied to DCEs. Models such as mixed logit, or hierarchical Bayesian models can also be used. These have the added flexibility of taking into account individual specific variables such as gender or other demographics.

Discrete choice experiments and conjoint analysis are not the only indirect elicitation methods that can be used to uncover preferences. It is also possible to observe people's actual behaviour in non-hypothetical settings or to use more psychologically grounded experimental methods. One such method, the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998) uses classification tasks to measure people's associations. For example, respondents are shown a sequence of African-Americans and white faces, and are asked to classify them as quickly as possible into a positive and negative category. The second task is similar but with the categories reversed: e.g. white faces into the positive category and African-American ones into the negative. Using reaction time measures as well as error rates, one can deduce what association is easier and more fluent for the participant, thus uncovering their implicit associations. The Implicit Association Test has been used in contexts as varied as social (Greenwald et al., 1998) and gender (Rudman & Kilianski, 2000) biases, implicit consumer preferences (Friese, Wänke, & Plessner, 2006) and political preferences (Arcuri, Castelli, Galdi, Zogmaister, & Amadori, 2008). Another indirect technique, the Affective Misattribution Procedure (B. K. Payne, Cheng, Govorun, & Stewart, 2005), uses priming to uncover implicit attitudes. A prime is displayed very briefly (usually around 100ms) before a neutral target (e.g. a Chinese pictogram) which respondents must evaluate as positive or negative. The theoretical backing of the method is that the prime subconsciously activates an either positive of negative reaction which affects the subsequent evaluation, so that the evaluation of the target is presumed to provide information about the degree of positive or negative associations of the prime. This procedure has been used to uncover racial bias (Ditonto, Lau, & Sears, 2013) and political preferences (Maier et al., 2015).

Although indirect elicitation methods have been designed as a way to overcome the weaknesses of direct methods, they have their own set of limitations. For instance, discrete choice experiments and other choice based methods rely on people's analysed choices being comparable with their choices in the real world. Yet in many cases choices in an experimental setting and are difficult to incentivise: either because the products do not exist or because of the costs associated (for example a DCE looking at the choices between cars). The difference between these choices is called hypothetical bias (List & Gallet, 2001). Although some studies claim that such limitations invalidate the use of DCEs, especially for non-market goods (Luchini & Watson, 2014), empirical studies have painted a more optimistic picture of the applicability of this technique. For example, Lusk and Schroeder (2009) compared a hypothetical and non-hypothetical choice experiment and concluded that the hypothetical bias affected the predicted probability of buying, as well as the willingness to pay for steaks. The differences in attribute importance, on the other hand, was not statistically significant, showing that the method is valid in eliciting preferences. These results have been found in other product groups (e.g. R. Moser, Raffaelli, & Notaro, 2014), thus justifying the widespread use of choice experiments in for preference elicitation and choice prediction (Lusk, Pruitt, & Norwood, 2006). There are other biases that effect DCEs such as the number-of-attribute and number-of-levels effects (Currim, Weinberg, & Wittink, 1981; Verlegh, Schifferstein, & Wittink, 2002). The number-of-attribute effect corresponds to the artificially increased preference that respondents have towards products with additional attributes. The number-of-levels effect occurs when respondents put more importance on an attribute that has significantly more levels than the others. For instance if there would be 10 types of milk but only 3 brands, there is a risk that the type of milk will artificially come out as the more important attribute.

A central tenet of choice based indirect elicitation methods is that a person's choice is reflective of their underlying preference. Yet even this is highly debated. For example, the centre stage effect (J. I. Shaw, Bergen, Brown, & Gallagher, 2000), whereby an item is preferred if it is displayed in the centre of a range can be used to argue against that. Indeed, if a person chooses a central item in a choice task, is it reflective of a true preference for the product itself, or for its positioning in the range. While introspecting has been shown to moderate this effect (Valenzuela & Raghubir, 2009), in practice it adds additional complexity to the preference elicitation process. Warmup tasks and deliberation have an effect on choices (J. Huber, Wittink, Fiedler, & Miller, 1993; Wilson et al., 1989), hence using them before a

choice based elicitation task may remove the centre stage effect, but can change the choices that respondents make. The representativeness of preferences by choices is also problematic in when considering the three preference reversal effects: the attraction effect (J. Huber, Payne, & Puto, 1982), the compromise effect (Simonson, 1989) and the similarity effect (Tversky, 1972). All three effects can be described by a change of preferences caused by the addition of a third alternative into a binary choice, where both alternatives (A and B) are equally preferred. In order to present these effects, it is helpful to think of a binary decision between two products, where only two attributes are considered: quality and economy (see Figure 1). Alternative A has a high utility (3) on the economy dimension, and a low utility (1) on the quality dimension. Alternative B is the reverse, with a high utility on the quality dimension and a low utility on the economy one. Assuming both attributes are equally important to a decision maker, alternatives A and B are also equally preferred since their total utilities are equal (4). The compromise effect involves the addition of a compromise option (C), placed between them (with utility 2 on both dimensions), which is then preferred. The similarity effect involves a third option (S) which is slightly better than B on one dimension but worst on the other. This increases preference for the dissimilar option A. Finally, the attraction effect posits that adding a dominated option (D) to the binary choice situation makes the dominating choice (A) more attractive. Although this last effect has been criticized as dependent on the stylized representation used in the experimental design (Frederick, Lee, & Baskin, 2014; for a reply see Simonson, 2014), the issue is identical as with the centre stage effect: are the contexts in which choices are made affect the choices themselves or do they affect people's underlying preferences. We will return to this theoretically important question, when discussing the nature of preferences. From a more practical standpoint, it is sufficient to note that if the context in which the preferences are elicited similar to the one in which the choice behaviour is made then these preferences will indeed reflect the drivers of people's behaviour.

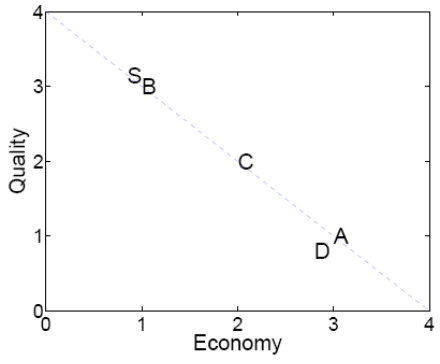


Figure 1: Example of a choice set illustrating the three preference reversal effects. While there are only options A and B in the set, they are equally preferred. The addition of S, C or D changes that preference in a deterministic way.

The non-choice based indirect methods previously presented, such as the implicit association test, are not without their own limitations. For example, they have been criticised for measuring mental associations rather than preferences. In addition to being more effortful to apply, they are also less granular than DCEs which reports preferences for each attribute of a product. Implicit methods usually measure a positive or negative association for a target product or a specific attribute (e.g. brand). Hence, although these can be used to help predict choices (e.g. Friese et al., 2006; Maier et al., 2015), they are more limited than full regression models outputted from discrete choice experiments.

Herein lies the strength of indirect preference elicitation: in addition to not relying on introspective access, methods such as conjoint analysis and discrete choice experiments can be used to identify the importance of the chosen product's attributes and to predict future choices. Although the empirical comparisons of predictive validity between direct and indirect methods has yielded mixed results (Sattler & Hensel-Börner, 2007), the latter are more extensively used and developed (Eggers & Sattler, 2011; Netzer et al., 2008). An interesting development for these methods is the inclusion of a direct elicitation step such as identifying unacceptable attribute levels before the choice task, or giving a direct initial rating of attributes (e.g. Green & Srinivasan, 1990; Voeth, Herbst, & Liess, 2013). Although respondents often make choices which violate these initial decisions (Green, Krieger, & Bansal, 1988), which itself can be seen as further evidence for limited introspection, taking them into account can improve the predictive abilities of resulting models (Voeth et al., 2013).

To sum up, direct elicitation methods are often easier to apply as they take less time and often require less effort from participants. This is especially true in complex environments with many attributes, in which these methods can be the only feasible ones. They are unfortunately also considered less precise in their measurements, and the often unconscious effect that the response scales have on the answers can reduce confidence in the results. Careful design of the elicitation task and the response scales can overcome some of these limitations. As these methods rely heavily on people having direct access to their preferences, their malleability is often used as evidence against people's reliable ability to introspect. Indirect elicitation methods, on the other hand, do not depend on introspective

access. But they are also affected by biases, such as positioning effects or the number of attributes effect. Additionally, the existence of preference reversal effects, which show that choices can be changed in a systematic manner, for example by adding non-preferred products to a decision context, question the nature and validity of the preferences that they elicit. Still, indirect elicitation methods are generally considered more reliable and are heavily used both commercially and academically, especially due to their ability to predict choices.

Choice models and prediction

One of the aims of preference elicitation is to be able to subsequently predict people's choices. Attribute importance measures are used to understand why people make decisions, while choice prediction often allows for external validity of these findings both within and outside of the laboratory. Prediction is also very important for marketers both for the optimisation of new products, and for market-share predictions of their current products. Experimentally, predictions are often validated using a subset of choices which have not been used in deriving the model (e.g. Elrod, Louviere, & Davey, 1992; Green & Srinivasan, 1990), called a holdout set. For example, if a choice experiment data consists of 300 choices, one can use 150 or 200 of these to create the model, while the rest are set aside as the holdout set for validation. Traditionally, choice prediction models assume that people make decisions 'as if' they were weighting and adding all the attributes of each considered alternative to get a total utility. As with the utility maximising model that is calculated from the DCE results, the alternative with the highest utility is chosen. Direct elicitation methods can also be used to parameterise these models, by simply mapping the attribute weights to the corresponding betas (G. P. Huber, 1974; Srinivasan, 1988). Such weighted additive models (WADD) have been successfully applied to a number of domains such as purchasing decisions (Netzer & Srinivasan, 2011) or job choice (Punj & Staelin, 1978).

Another class of models, the heuristic based approach (Gigerenzer, Todd, & the ABC Research Group, 1999), involves a more psychological realistic assumptions and takes into account the cognitive limitations faced by decision makers (Gigerenzer & Goldstein, 1996). Models such as Elimination by aspects (Tversky, 1972) or the lexicographic decision rule (Bettman, 1979) do not require an evaluation of all the attributes of each product in the choice set. Instead, they evaluate the attributes by order of importance. Elimination by aspects, for example, uses successively all attribute levels from the most to least preferred,

and eliminates alternatives from the choice set who do not correspond to these preferences. As each attribute level, or cue, is evaluated the choice set is reduced; this implies that fewer cognitive resources are needed in order to make the choice, as compared to WADD models. These models can be easily parameterised by both direct and indirect elicitation methods, as all they require is an ordering of the importance of the attributes and their levels[2]. In some cases, it is even possible to infer the lexicographic ordering of attributes from conjoint analysis or choice experiment data (Kohli & Jedidi, 2007; Yee, Dahan, Hauser, & Orlin, 2007), but models parameterised this way are less predictive than their WADD counterparts. Otherwise, the attribute importance can also be taken from the weights in the corresponding WADD model, which adds an extra step parameterising the lexicographic model.

There is evidence that people use heuristic based models when making decisions under time constraints (Dhar & Nowlis, 1999), where information is costly (Bettman, Luce, & Payne, 1998; Gigerenzer et al., 1999; J. W. Payne, Bettman, & Johnson, 1993) or when the number of attributes or alternatives is large (Einhorn, 1971; Westenberg & Koele, 1994). Hybrid approaches promote the use of lexicographic models to reduce the number of alternatives in a consideration set (Hauser, 2014), before the final choice is made using a WADD model (Gilbride & Allenby, 2004). Still, in general it is still WADD models which are used for prediction. This is due to their ability to take into account individual characteristics (such as age or gender) and their usefulness in estimating market shares.

The nature of preferences

The elicitation methods and the choice prediction models that are described in the previous section bring to light an important discussion with regard to the nature of preferences. Indeed, some models rely on preferences having numerical values, which can be used as weights in an expected utility model, while others only require a relative ordering of attributes. These theoretical considerations have an important impact on elicitation, as wrongly applied scales may lead to incorrectly elicited values. If the attribute preferences are relative, attempting to uncover them by asking for an actual value on a numerical, metric scale, may lead to inconsistencies. In order to be usable for prediction purposes, theories of choice need to be able to account for how people make choices, including the three main

[2] Some models require an ordering of the attributes (e.g. type, size, brand…), and within each attribute the levels, while others treat each attribute level (e.g. organic, 2 pints, 4 pints, Supermarket brand…) as a distinct feature and require only an ordering of those.

preference reversal effects noted earlier (the attraction, compromise and similarity effects). Some theories of choices and preferences, such as expected utility theory (Von Neumann & Morgenstern, 1944), regard preferences as being value based and often assume the independence of irrelevant alternatives (IIA). This in turn leads to the assumption of a stable utility or preference for each alternative. The IIA axiom also states that the utility of an alternative is not dependent on the other alternatives in the choice set, an alternative will be preferred within a choice set regardless of all additional options. This is clearly violated by the preference reversal effects, whereby the addition of a strategically constructed new option in a choice set can change which is the preferred alternative. Extensions to value based theories, such as the third generation prospect theory (Schmidt, Starmer, & Sugden, 2008), have made it possible to explain these effect, for example by adding uncertain reference points. Non-value based models of choice are often reliant upon comparisons and therefore acknowledge from the outset the effect of the context on decisions. Elimination by aspects, for example, compares the attributes of each alternative to the elicited ordering of preferences, but others also entail comparisons between alternatives (Stewart et al., 2006). Comparative models therefore make less assumptions about the nature of preferences, requiring only the ability to make binary comparisons, and are often seen as more psychologically plausible (Vlaev, Chater, Stewart, & Brown, 2011).

Dual preference framework

Up to this point, we have been assuming that there is one type of preference which is then used by the different models to make choices. A contrasting view, is that there are in fact, two different types of attitudes or preferences: an implicit preference which is usually not directly accessible and can be uncovered through indirect elicitation methods (Greenwald & Banaji, 1995; Greenwald et al., 1998) and an explicit one which is usually easily reported by direct methods. Implicit preferences are seen as less susceptible to effects such as social desirability and are less malleable than explicit ones (Gregg, Seibt, & Banaji, 2006). Social behaviour, such as attitudes, self-esteem and stereotypes are seen to have an important implicit component (Greenwald & Banaji, 1995). For example, the halo effect (Thorndike, 1920) can be seen as evidence for an implicit component in attitudes. This effect is similar to anchoring, in that the evaluation of a target attribute is affected by another attribute. Nisbett and Wilson (1977b) demonstrated this effect by showing that ratings of an instructor's mannerisms and accent were positive when he was acting in a warm and friendly way, and

negative when he was cold and distant. Thus, the evaluation of the instructor's attributes (e.g. accent) were affected by his overall evaluation. Participants did not report any introspective access to this effect, which can also be taken as evidence for the limited access people have to at least part of their preferences. The existence of implicit attitudes within a dual attitude framework has also been used to predict self-reported undecided people's preferences outside of the laboratory (Arcuri et al., 2008; Galdi, Arcuri, & Gawronski, 2008). Hence, it may be that implicit preferences are precursors to explicit ones, or that they can affect the creation of explicit preferences. In other cases, explicit and implicit preferences can be in opposition. In these instances, implicit preferences are more predictive of choices under high cognitive load (e.g. time pressure) while explicit preferences are more predictive of more deliberative choices (Friese et al., 2006). When presenting the various direct preference elicitation techniques, we have noted how direct elicitation methods, which by definition measure explicit attitudes and preferences, are affected by the nature of the response scales that are used and the question order. Implicit attitudes, which are less malleable, are therefore used to stabilise the overall attitude (Gregg et al., 2006).

This dual attitudes framework is not without criticism. Often, the existence of implicit, inaccessible preferences is identified by the difference between explicit and implicit elicitation methods. For example, Gawronski, LeBel and Peters (2007) refute this stance by challenging the three common assumption of implicit preferences: their inaccessibility, their resistance to social desirability and their stability. They argue that these differences, between explicit and implicit preferences, are due to measurement errors and lack of conceptual correspondence between the measures. For example, once corrected for measurement error, self-reports are more correlated with indirect preference elicitation measures (Gawronski, 2002). Similarly, both techniques often differ in the specificity of their target. Hence, the dissimilarity between the Modern Racism Scale (McConahay, 1986) and the race-IAT (Greenwald et al., 1998) is due to the former measuring attitudes towards political issues and discrimination while the latter measures responses towards individual members of the group. The stability of implicit attitudes and their resilience to motivational factors such as social desirability has also been challenged. As mental imagery (Blair, Ma, & Lenton, 2001), exemplars (Dasgupta & Greenwald, 2001) as well as changes in the context of the stimuli (Wittenbrink, Judd, & Park, 2001) all affect implicit attitudes, it appears that such attitudes are not exempt from both deliberative and non-deliberative changes. Hence, Gawronski et al. (2007) posit that there are not two different types of attitudes. Instead, implicit methods only

measure activation, while explicit preferences reflect the outcomes of a validation process. The existence of such a process therefore raises an additional question with regard to the nature of preferences, or more specifically the nature of the process from which they arise, as will be discussed in this next section.

Preference construction

The final important point on the nature of preferences that I wish to touch upon is whether they exist as stable constructs and are therefore only retrieved from memory, or if they are constructed when they are needed. The context dependence of preferences, exemplified by the preference reversal effects or the question order effects discussed previously, is often presented as evidence for the constructive nature of preferences. In this view (Bettman et al., 1998), people do not have a set of stable preferences that they access when needed, rather they use the task at hand, as well as the available information, to construct their preferences. This framework also acknowledges that preferences are not the outcome of a single invariant process such as an expected utility WADD model. Rather, taking into account the notion of bounded rationality (Simon, 1955), preferences are the outcome of different processes depending on the context. For example, in a context where there is very little time to make a decision, the observed preference is likely to be the output of a simplifying heuristic such as elimination by aspects, rather than a weighted additive process, which is usually present in more deliberative tasks. Hence, in such a constructivist view people's choices are dependent on a number of external and internal factors. Internal factors can be goals, for example decision makers may decide that it is more important to reduce cognitive load or to maximise the ease of justification in certain circumstances. External factors include the other alternatives available and the framing of the decision task. A more specific example of framing is provided by the effect of loss aversion (Kahneman & Tversky, 1984), which refers to the asymmetry in valuation of losses and gains. Kahneman and Tversky (1984) give the example of framing a choice in terms of gambles. When participants were told to assume that they are richer by $300 and given a forced choice between a sure gain of $100 or a 50-50 per cent chance to gain $200 or gain nothing they overwhelmingly chose the sure gain. While when they were told to assume they were richer by $500 and given a forced choice between a sure loss of $100 or a 50-50 per cent chance to lose $200 or lose nothing, most participants expressed preference for the 50 per cent gamble of losing nothing. Hence the participants showed different preferences for risk depending if it

was framed as a sure gain or a sure loss, even though in both cases the choice was between $400 and an even chance of $500 or $300. If people's preferences are constructed and affected by so many different factors, it can be seen surprising that economical markets function at all. Interestingly, although initial valuations have a very important arbitrary component, subsequent valuations and decisions are coherent (Ariely et al., 2003). For instance, although people value wines differently if they have been subjected to a high or low numerical anchor, their subsequent valuation of other bottles of wine of different qualities are higher (lower) for better (worst) quality wines.

In addition to explaining the coherence of valuation markets, this also sheds a light on some of the mechanisms of preference construction, namely relative processing. It also justifies the use of comparative elicitation techniques such as anchored scales, which allows people to capitalise on the ease of comparative valuations compared to a specific scale. The malleability of initial valuations has been made even more salient by a number of experiments which showed that the same experience can be rated as positive or negative depending on the context around it (Ariely, Loewenstein, & Prelec, 2006). More specifically, Ariely et al. (2006) divided a class of undergraduate students into two groups. One group was asked if they would accept US$ 2 to attend a poetry recital by their professor, while the other half were asked if they would pay the same amount for the recital. In order to test the actual value assigned to the experience, both groups were then informed that the recital will be free and were asked their intention of attending it. The results showed clearly that the group which was going to be paid to attend rated the experience as less pleasurable, as expressed by a very low number of students intending to attend the recital for free. On the contrary, the group who was asked if they would pay to attend, majoritarily answered that they would attend for free. Most surprisingly, these results held even when participants were given a sample of the poetry recital experience and were made clearly aware of both conditions: students who had to think about how much money they would pay for the experience rated it as more pleasurable than those who were asked if they would be paid. Thus, even when given the opportunity to sample the experience, preferences were constructed using the context. The experience studied was ambiguous on purpose, as there are some experiences or products which can be valued as positive or negative more easily. Still, there is evidence that even when such an experience is sampled people do not have a fixed preference, but instead construct one using contextual cues such as numerical anchors or being paid rather than paying for the experience. Anchoring does not only affect valuations and preferences, but

also recollection of one's own, and judgement of other people's, behaviour. For example, when completing paper and pencil mazes, or observing someone complete these mazes, participants were just as susceptible to be affected by anchors in reporting how many were completed (Cheek et al., 2015). This was not limited to cognitive tasks, as their estimation of how many steps they have just climbed was also affected by such anchors.

Opponents of the aforementioned constructed preference view argue for the existence of stable inherent preferences, which are not influenced by context effects. This view acknowledges that people find it difficult to make absolute valuations, and that this causes the inconsistencies discussed above, but states the existence and importance of inherent preferences. Simonson(2008b) defends this point of view by noting that many of the context sensitive preferences are transient as opposed to enduring ones which influence subsequent behaviour. He adds that the evidence cited for the construction of preferences is also very much dependent on the elicitation method provided, and that in non-artificial settings these experimental effects are overstated. This is echoed by findings which suggest that the attraction effect is dependent on how the alternatives are represented (e.g. visually vs numerically) or if they are experienced (Frederick et al., 2014) and may therefore not be as widespread in the real world as previously thought. Simonson (2008b) states that inherent preferences are uncovered for example by experience. According to this viewpoint, the adoption of new technologies with novel features, such as the iPhone with its lack of physical keyboard, or Nintendo's Wii with its motion-sensitive remote, reveal that people had inherent preferences for these novel features. Inherent preferences are also presented as more similar to the implicit preferences that we previously discussed in that they are less resistant to change and and spontaneously retrieved. Simonson's paper has started a flurry of discussions (e.g. Bettman, Luce, & Payne, 2008; Kivetz, Netzer, & Schrift, 2008; Simonson, 2008a) surrounding the existence and conceptualisation of these stable inherent preferences. Smith (2008) for example refutes the usefulness of this definition of inherent preferences, opting instead for dual attitudes and implicit associations theories as better constructs. Unfortunately, the existence of stable inherent preferences is still very difficult to falsify (Bettman et al., 2008) but remains pragmatically important to the study of preferences as neither can it be dismissed (Simonson, 2008a). This discussion on the nature of preferences has important implication for preference elicitation. Indeed, constructed preferences should be measured differently to preferences which just have to be accessed (J. W. Payne, Bettman,

& Schkade, 1999). In general, indirect preference elicitation methods such as DCE take this into account by mirroring as closely as possible the context in which decisions are made.

Aims of the studies

The three studies that will be presented in the continuation of this thesis each have different yet complementary aims, which together expand our understanding of the limits of introspection.

The first study has two specific aims:

- Aim 1: Examine the consistency of the ranking of attribute importance of familiar products collected by 3 direct elicitation methods, in a sample of UK consumers.
- Aim 2: Compare the accuracy of the 3 direct elicitation methods and an indirect elicitation method (DCE) to predict choices of these familiar purchases in terms of hit-rates (% of correct predictions) by using two modelling frameworks (LEX and WADD).

Aim 1 checks the internal consistency between different introspection based methods, while aim 2 compares the validity of introspective and non-introspective methods on choice prediction. In order to reach these aims, I designed an experiment which elicited preferences for familiar products using a number of direct and an indirect method. I then compared the attribute importance rankings between the methods, following which I constructed predictive models for each elicitation method, both in a lexicographic and a weighted additive framework. I finally compared the hit-rates of each model on a sample of experimental choices as well as a sample of real world ones derived from a large supermarket scanner dataset.

The second study built upon the first one, in using the same methodology and product classes but the aims were different.

- Aim 1: To examine the consistency of direct preferences for ethical, occasionally ethical and non-ethical consumers.
- Aim 2: To explore the degree to which direct and indirect preferences converge for ethical, occasionally ethical and non-ethical consumers.

Both these aims extend the first study, by eliciting and comparing the preferences of three distinct groups of consumers. Given the explicit importance that ethically minded consumers give to ethical attributes it was expected that such an analysis would uncover strong differences between groups in terms of direct as well as indirect preferences. To achieve this aim, I grouped individuals form a sample of UK consumers into three groups dependent on how often they reported buying ethical products: always, occasionally/frequently or never. I described these three groups in terms of their demographic and attitudinal characteristics, ensuring the internal coherence of the groups. I then compared the attribute importance rankings provided by the direct elicitation methods between groups. The indirectly elicited attribute importance orders were also compared, before examining the intra- and inter-group preference heterogeneity.

The third study extended this work into the domain of political attitudes with the following two specific aims:

- Aim 1: To verify the agreement between directly and indirectly elicited political attitudes
- Aim 2: To check if directly and indirectly elicited political attitudes show any short-term change after exposure to positive advertisement

In order to reach these aims, I randomly divided a sample of urban Pakistani television viewers into three conditions, thus informing which advertisement they would see: a neutral Pakistani advertisement for washing liquid, an informationally rich pro-american advertisement and an informationally poor one. The direct and indirect attitudes for a number of countries, including the USA were elicited, using door to door canvassing as part of a TV ratings panel. One direct elicitation was used before showing them the advertisement and the other elicitation methods were completed afterwards. For the analysis, I compared the directly elicited attitudes before and after the advertisement (within conditions). I also compared the post-advertisement directly elicited attitudes between conditions, as well as the indirect elicited ones. I finally compared the elicited attitudes to similar ones from professional surveys in Pakistan, to provide external validity to the findings.

Chapter 3. The limits of introspection: the inaccessibility of preferences even for familiar products.

Abstract.

The rise in popularity of indirect preference elicitation methods such as conjoint analysis and discrete choice experiments represents the lack of confidence in people's ability to introspect and express their preferences. The comparison between direct (introspection-based) and indirect methods has yielded mixed results. In this paper, I propose to research the limits of introspection, specifically for commonly purchased products for which preferences should be more accessible. In doing so, I compare the consistency of the ranking of attribute importance for these familiar products, as elicited through three direct elicitation methods. I also compare the direct and indirect elicitation methods in their ability to predict choices.

Before comparing the methods amongst themselves, I make sure that the choices made experimentally mirrored the real world behaviour, by comparing experimental market-shares with ones extracted from a large supermarket scanner data set. Following this validation, the different elicitation methods are compared on their resulting attribute importance scores and their ability to predict choices through weighted additive and lexicographic models. The results provide evidence for a lack of reliable introspective access, thus justifying the widespread use of choice experiments. Interestingly, in this context of habitual decisions and direct preference elicitation, lexicographic models are shown to dominate their weighted additive counterparts thus providing a new framework for directly elicited preferences and choice prediction.

Introduction

One of the central assumptions underlying surveys and other direct preference elicitation methods is that people have introspective access to their preferences and can communicate them. The aims of preference elicitation are usually twofold: extracting attribute importance measures and predicting future choices. Attribute importance measures allow for a better understanding of what features of a product consumers value, but predicting choices enables an external validation of these identified intrinsic motivations. Views of the nature of preferences, which affects elicitation methods, can be divided into two groups. The

inherent preference view (Simonson, 2008b) accepts the existence of stable underlying preferences which guide behaviour, and explains elicitation errors or inconsistencies as biases in the reporting and accessing processes. On the other hand, the constructed preference view (Lichtenstein & Slovic, 2006) suggests that people construct valuations at time of elicitation using cues from the wider environment. Taking a more nuanced position, Fischoff et al. (1988) state that "People are most likely to have clear preferences regarding issues that are familiar, simple and directly experienced". These conflicting views offer additional challenges to the study of introspection, as the direct elicitation of non predictive preferences could be due to either limits of the introspective process itself or the instability of the elicited preferences.

It is important to note that research into preferences, and preference measurement, often use products that are not very commonly purchased: televisions, calculators, cars, etc., and hence where people's knowledge of the products and their own preferences may be relatively slight. For example, Sattler and Henserl-Börner (2007) reviewed 23 studies and only 2 of those involve regularly purchased products (in both cases, coffee). In this paper, I explore the limits of people's introspection of their preferences for commonly purchased grocery products, using attribute importance measures as well as predictive models. In order to externally validate my experimental results, I make use of a large supermarket scanner dataset, which also ensures the familiarity of the products used. These often experienced products should have clearer preferences, if such a construct exists. Additionally, since the products are brought nearly every week and therefore more salient, it is more plausible that people's actual purchases will be reflected experimentally even though they might not be able to report their preferences. The aim of this paper is twofold: on the one hand, the comparison of the rankings of attribute importance for familiar products collected by different direct elicitation methods will inform us of the stability of directly elicited preferences for these products. On the other hand, the comparison of the accuracy of direct and indirect preferences in predicting choices will provide an additional measure for the reliability of these preferences in predicting choices. I also put forward the applicability of a directly elicited lexicographical methodology to predict choice amongst a complex range of products. The results partially support this new methodology but point towards the unreliability of direct preference elicitation techniques, and therefore support the conclusion of limited introspective access even for habitual decisions.

The rest of the paper is organized as follows: in the next section, I give more details about the different theories regarding preferences, and present the preference elicitation methods used as well as the predictive models they are validated with. After presenting the theoretical background, I describe the experimental method in detail, followed by the results which suggest a lack of reliable introspective access, concluding with a general discussion.

Literature review

Limited introspection

Marketing, among other disciplines, often relies on direct querying of people's preferences and behaviours using interviews or surveys. The assumption that people can introspect on their internal states has been long debated in psychology (Nisbett & Wilson, 1977a), and causes concern as to the validity of the elicited results. Examples of lack of introspective access range from misinterpreting one's internal states (Cantor, Zillmann, & Bryant, 1975; Dutton & Aron, 1974; Schachter & Singer, 1962), to biases in reporting on one's own behaviour (Schwarz & Oyserman, 2001). Response mode effects such as framing and anchoring (Ariely et al., 2003; Cheek et al., 2015; Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991; Tversky & Kahneman, 1974) have also been put forward as evidence for lack of introspective access. Indeed, if people's self-reports are so dependent on how the questions are posed, or the order in which they are asked this might be indicative of a lack of self-knowledge or even the lack of stable preferences underlying behaviour (Bettman et al., 1998). Supporting this view, studies such as Ariely et al's (2006) show that context can influence people's valuation of an experience, even when sampling the experience and therefore having an opportunity to construct a preference value prior to elicitation.

Preference elicitation

In the context of purchases, introspective access is usually explored using preference elicitation techniques. These methods can be divided into direct and indirect methods. Direct methods, which rely on introspective access, often require respondents to express their attribute preference directly, by ranking, sorting (e.g. Q-sort: Stephenson, 1953) or rating the attributes and their levels (e.g. Magnitude scaling: Lodge, 1981). The most widely used of these techniques is the simple metric scale. Since respondents are not constrained in how to rate attributes, responses tend to be grouped at the high end of the scale (Meyer, 1999). In an attempt to improve this measure, a number of similar direct, self-explicated, methods have been proposed. The Magnitude Scaling technique (Lodge, 1981), provides an anchor thus

making the scale more explicit. Building on this, Huber's (1974) client-explicated parameter estimates method uses two anchors to elicit attribute importance: the least and most important attributes serves respectively as the minimum and maximum anchors against which the other attributes are rated. Finally, to increase reliability, Srinivasan's (1988) self-explicated approach eliminates unacceptable attribute levels before following Huber's methodology. Another way of improving the simple metric scale is by forcing trade-offs between attributes, as is exemplified by the commonly used constant sum scales (CSS). This method requires respondents to distribute 100 points among the attributes, in line with how important they perceive them.

Direct elicitation methods depend strongly on introspective access and any limitations in that process will therefore be strongly mirrored in the results. For example, it has been observed that participants' choices can involve attributes that they previously rate as 'unacceptable' (Green et al., 1988). Direct methods have a number of other possible problems (Green & Srinivasan, 1990): they are affected by common survey issues such as question order effects (Schuman & Presser, 1981; Schwarz, 1999), response scale effects (Schwarz et al., 1991), and social desirability bias (DeMaio, 1984). The advantages of these methods is that they are easy to design, faster to collect, usually less straining on respondents, and associated with lower costs especially in studies with large number of attributes (Sattler & Hensel-Börner, 2007).

Indirect preference elicitation methods, which have become the most used preference elicitation techniques despite their added complexity, deduce the importance of the product attributes by analysing choices or ranking of products. In this way they circumvent the issue of introspection by relying on externally analysing choices and behaviour rather than requiring participants to do so themselves. The validity of this approach rests on the assumption that hypothetical choices in an experimental setting are comparable to real decisions. The difference between these decisions is known as a hypothetical bias (List & Gallet, 2001; Little & Berrens, 2003; Murphy, Allen, Stevens, & Weatherhead, 2005) and few studies have investigated this in the field (R. Moser et al., 2014). The findings in general point towards the existence of such a bias, leading for example to over-estimation of willingness to pay (Lusk & Schroeder, 2004) values or inconsistent performance of predictive models (Hudson, Gallardo, & Hanson, 2012). The most popular indirect preference elicitation methods are conjoint analysis (Green & Rao, 1971) and discrete choice experiments (DCE; Wittink et al., 1994), which require participants to repetitively make choices (or rating) from

different ranges of products. In order to limit the number of choices and trade-offs that must be analysed, many conjoint variants include a direct elicitation section (Eggers & Sattler, 2011; Green & Srinivasan, 1990; Voeth et al., 2013) where participants are asked to identify 'unacceptable' levels, which in turn helps to inform the composition of the products to be compared.

Most indirect methods are more time consuming than the direct ones, as they require a large number of choices to be made by respondents. Additionally, these methods also suffer from their own biases such as the number-of-attributes effect and the number-of-levels effect (Currim et al., 1981; Verlegh et al., 2002). These effects artificially increase the preference for products having more attributes or increase the importance of a particular attribute with additional levels. A balanced experimental design can help circumvent these issues.

External validation of preference elicitation methods is usually conducted not only by comparing the measured attributes importance, but also by predicting choices (see

Figure 2). Thus, models of choice are parameterised using elicited preferences and are then compared in terms of predictive ability. There are a number of possible models that can be parameterized from each elicitation methods. For example, indirect methods such as DCE are used to derive multinomial logistic regression models that are then used to predict choices or market-shares of products.

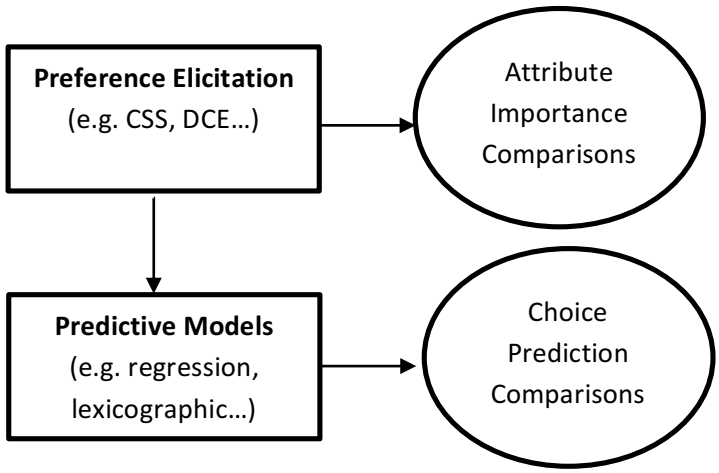


Figure 2: Flow of the comparison between preference elicitation techniques. The measured preferences themselves can be compared in terms of attribute importance, or they can be used in models to predict choices.

Modelling frameworks

Traditional predictive models of multi-attribute decision making assume that consumers make decisions 'as if' they were weighting all the attributes of each alternative available and adding them up to get a general preference value or utility, choosing in each situation the alternative with the highest such value. The preference weight of each attribute is directly extracted from a direct preference elicitation task, or deduced from an indirect one. The resulting weighted additive models (WADD) have been successfully used to investigate and predict decisions in many different contexts including healthcare (De Bekker-Grob, Ryan, & Gerard, 2012), purchasing decisions (Hudson et al., 2012; Netzer & Srinivasan, 2011), and job choice (Punj & Staelin, 1978). Using directly elicited weights, such models are trivial to parameterise. Most self-reported attribute importance values are directly equivalent to the parameter weights in the WADD model (G. P. Huber, 1974; Srinivasan, 1988). The elicited values in a CSS or a metric scale can be rescaled or used as is in a prediction model. On the other hand, valueless elicitation methods such as Q-sort or ranking cannot be easily used in these models. The lack of an interval or ratio scale makes the interpretation of the distance between two attributes difficult as it is impossible to quantify the difference between each value. For indirect methods, such as DCEs, the parameters are estimated using multinomial logistic regressions (Louviere et al., 2000).

In contrast, the research tradition of simple heuristics (Gigerenzer et al., 1999) proposes a number of more psychologically realistic decision models which take into account the cognitive limitations faced by decision makers (Gigerenzer & Goldstein, 1996), and are therefore better adapted to make decisions under attention, time and memory limits. The lexicographic semi-order model (LEX; Tversky, 1969), is a good example of such a model. Rather than calculate a value for all the available products, the model assumes that the features and feature levels have a preference order. When choosing between a number of products, the model selects the product with the most favoured feature. If this results in more than one product being chosen, the next most important feature is used to discriminate between them, and so on. As they process the attributes in a specific order, making a decision as soon as one attribute has managed to differentiate a single option from the choice set, LEX models use on average less information than WADD models which integrate all the attribute values, while leading to less effortful and faster decisions. There is evidence that decisions makers use this type of strategy when making decisions under time constraints (Dhar & Nowlis, 1999), or where information is costly or effortful (Bettman et al., 1998; Gigerenzer et

al., 1999; J. W. Payne et al., 1993), or when the number of attributes or alternatives is large (Einhorn, 1971; Westenberg & Koele, 1994). Additionally, in multi-stage "consider-then-choose" models, similar heuristics are used in predicting consideration sets (Bettman & Park, 1980; Gilbride & Allenby, 2004; Hauser, 2014) while WADD models are then used on these smaller consideration set to make the final decision.

In terms of preference elicitation, the lexicographical framework could be argued to be more psychologically plausible since it does not assume that the brain computes or stores absolute values (Vlaev et al., 2011), and it processes information in a sequential manner (Martignon & Hoffrage, 1999). In this view, only the relative ordering of the different attributes is important, which also minimises the sensitivity of the model to the underlying elicitation techniques. Yet, in practice WADD models are still more widely used.

Comparison metrics

Preference elicitation methods are often compared on two metrics: attribute importance (Chrzan & Golovashkina, 2006; Louviere & Islam, 2008; Schulz, Speekenbrink, & Shanks, 2014; Srivastava, Connolly, & Beach, 1995) and predictive validity of the resulting models (Dieckmann, Dippold, & Dietrich, 2009; J. Huber et al., 1993; Jain, Acito, Malhotra, & Mahajan, 1979; Sattler & Hensel-Börner, 2007; Scheibehenne, Miesler, & Todd, 2007). Attribute importance comparisons between preference elicitation methods show relatively high agreement within direct and indirect methods, but low agreement between the two categories (Louviere & Islam, 2008). While attribute importance scores inform us on the internal consistency amongst methods, the external validity is provided by the predictive measure. Predictive validity can be compared both experimentally, by comparing model prediction to choices made in a discrete choice experiment for example, and in the field. A critique voiced against extrapolating general behavioural insights from experiments is that they do not always reflect people's behaviour in the real world (Levitt & List, 2007; Voors, Turley, Kontoleon, Bulte, & List, 2012). This is also known as hypothetical bias: the notion that hypothetical choices do not reflect actual ones. In this case, for DCEs, studies have shown that hypothetical bias is minimal and that choices made in a discrete choice experiment are generally comparable with those made in the field (Adamowicz, Boxall, Williams, & Louviere, 1998; Lusk et al., 2006; Lusk & Schroeder, 2004). The similarity can be reinforced using incentive alignment techniques (Ding, 2007; Ding, Grewal, & Liechty,

2005; Netzer et al., 2008) and to some extent using a "cheap talk" script (Cummings & Taylor, 1999; Lusk, 2003; R. Moser et al., 2014). As the experiment uses neither, I compare choices made experimentally with actual sales data as an additional measure of external validity.

Comparisons of predictive validity between direct and indirect preference elicitation methods have shown mixed results. Although proponents of indirect methods praise their superior predictions, empirical comparisons with direct methods have failed to robustly support this (Green & Srinivasan, 1990; Sattler & Hensel-Börner, 2007). Surprisingly, this is even true for some more advanced hybrid methods (Netzer & Srinivasan, 2011; Srinivasan & Park, 1997). In a review of comparative studies (see Table 1 for an extended summary) between various methods, Sattler and Hensel-Bröner (2007) found that not only was there no strong evidence for one approach over the other, but only a small minority of studies were conducted on products or services which the participants had a lot of experience choosing between. Since the familiarity, even within a choice experiment, can change the performance and reliability of preference elicitation methods (J. Huber et al., 1993) it is important to specify a suitable decision context for the tasks.

Decision context

Taking into consideration that choices for familiar products are more likely to be associated with stable preferences (Fischhoff et al., 1988), I explore introspection through preferences for supermarket grocery shopping products and apply the findings to a large supermarket scanner dataset. Although the products purchased are often familiar and the decision context habitual, inconsistencies have been found between what people express is guiding their decisions and what is revealed by their behaviour. For example, while price is often stated as an important factor guiding consumers' decisions, Dickson and Sawyer (1990) observed that supermarket shoppers spent very little time deliberating among products and were unable to recall the price of the selected item. This short deliberation time could be indicative of a heuristic decision process (Gigerenzer et al., 1999), whereby consumers do not use all the available information. Yet, this contrasts with findings that when asked directly, supermarket consumers tend to request additional information and labels on the products (Gadema & Oglethorpe, 2011) rather than acknowledge that they use limited information. Are consumers really aware of how much information they use? Although direct elicitation of preferences have had some success in predicting choices (Srinivasan, 1988; Srinivasan & Park, 1997) and consideration sets (Ding et al., 2011), indirect or hybrid methods (Eggers &

Sattler, 2011; J. Huber et al., 1993; Netzer et al., 2008) have mostly superseded them as the industry standard for preference measurement. Specifically, these methods can capture influences which people are unaware of. For example, Muller, Lockshin and Louviere (2009) showed differences in packaging importance between stated and revealed preference elicitation methods. In their study, respondents rated as relatively unimportant visual cues such as packaging, when asked about their wine preference using a direct elicitation method. An analysis of the same respondents' choices in a DCE revealed that these visual attributes had a much more important effect on choice outcome.

Generally, comparisons between direct and indirect elicitation methods have used non-habitual decisions such as cars or apartments (see Table 1). This study specifically targets familiar decisions under the assumption that introspective access to them will be greater. The familiarity of the decisions is not limited to the product classes under consideration, but also to the products themselves which are directly extracted from the supermarket data, thus increasing the likelihood that participants would have seen the exact same products before, during their grocery shopping. The current study also adds a more complete comparison between elicitation methods, both in terms of the elicitation methods tested (three direct elicitation techniques and an indirect one) and the modelling frameworks that are used for prediction purposes. Although the models are aggregated across participants, validation is carried out on both hypothetical choices, as in much of the existing literature, but also on actual supermarket scanner data.

Source	Methods Compared	Choice prediction	Prediction range size	Results a)	Products	No of Attrs.	Sample Size
Srinivasan (1988)	Conjoint trade-off — SEM	Actual choices	2 or more	Non-significant differences	Job offers	8	54
Green, Carmone and Wind (1972)	Conjoint — SEM	Hypothetical	27	No major differences b)	Discount cards	3	43
Leigh, MacKay and Summers (1984)	Conjoint — SEM	Actual raffle choices	10	Non-significant differences	Pocket calculators	5	122 e)
Green and Helsen (1989)	Conjoint — SEM	Hypothetical	8	Non-significant differences	Apartments	6	99 e)
Huber et al. (1993)	Conjoint — SEM	Hypothetical	2-3	Non-significant differences	Refrigerators	5 and 9	393
Green, Krieger and Agarwal (1993)	Conjoint/ACA — SEM	Hypothetical	12	Mixed results for different measures. SEM better than ACA d)	Cars	8	133 e)
Huber, Daneshgar and Ford (1971)	Conjoint — SEM	Actual choices	15-20	SEM better than conjoint c)	Job offers	5	30 e)
Wright and Kriewall (1980)	Conjoint — SEM	Actual (multiple) choices	19-22	SEM better than conjoint	College applications	5	120
Green, Goldberg and Wiley (1982)	Conjoint — SEM	Hypothetical	4	Conjoint better than SEM	Household appliance	7	476
Akaah and Korgaonkar (1983)	Conjoint – SEM	Hypothetical	6	Conjoint better than SEM	HMO	6	80
Akaah and Korgaonkar (1983)	Huber/Green hybrid – SEM	Hypothetical	6	Non-significant differences	HMO	6	80

Green, Goldberg and Wiley (1982)	Green hybrid – SEM	Hypothetical choices	4	Green hybrid better than SEM	Household appliance	7	467
Agarwal and Green (1991)	ACA - SEM	Hypothetical	16-18	SEM better than ACA d)	Apartments	6	170 e)
Huber et al. (1993)	ACA - SEM	Hypothetical	2-3	ACA better than SEM	Refrigerators	5 and 9	393
Hensel-Börner and Sattler (1999)	ACA - SEM	Hypothetical	X	Non-significant differences	Coffee	8	144
Srinivasan and Park (1997)	Customized conjoint — SEM	Actual choices	2 - 3	Non-significant differences	Job offers	8	121 e)
Hensel-Börner and Sattler (1999)	Customized conjoint — SEM	Market shares	X	Customized conjoint partly better than SEM	Coffee	8	144
Dieckmann, Dippold and Dietrich (2009)	Conjoint — Revealed LEX	Hypothetical ranking and rating	2	Conjoint better	Ski Jackets	10	142 e)
Hauser, Dong and Ding (2014)	Conjoint — SEM — unstructured	Hypothetical	30	Mixed results. b)	Cars and Mobile phones	53 and 22	204 and 143 e)
Netzer and Srinivasan (2011)	Adaptive SEM — SEM — ACA	Hypothetical	2	ASEM best. No statistical difference between SEM and ACA	Digital Cameras	41	154 e)
This study	DCE – SEM (using WADD models and LEX orderings)	Hypothetical choices and actual sales prediction.	9 - 36	Mixed results, but SEM method less reliable in general	Fresh milk, tinned tuna, and eggs	26, 26 and 19	900-1000

Table 1: Empirical studies[3] comparing direct (SEM: self-explicated) and indirect (conjoint analysis, ACA: adaptive conjoint analysis) preference elicitation methods.

a) Significant differences between methods b) Significance not tested c) For experienced respondents; non-significant differences for non-experienced respondents d) Exception: ACA better than the Self- explicated part of ACA (Green, Krieger and Agarwal 1991) e) Respondents were students

[3] Adapted and extended from Sattler & Hensel-Börner (2007)

Method

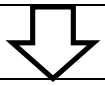
The Participants

Participants were recruited on an online commercial panel provider (Belindi) and consisted of a demographic mix of U.K. residents screened to be fluent English speakers older than 18 years old who frequently buy milk, tinned tuna and eggs. The median age was 52 years and 39% of respondents were male. Performance in the experiment was not incentivized, and participants were urged to give their true preferences in return for points in a loyalty program. Answers that were deemed inconsistent were excluded and not rewarded. Inconsistency was measured by asking participants for their year of birth at the beginning of the survey and their age at the end, answers that differed by more than 2 years were inconsistent. Additionally, participants that took less than 10 minutes to complete the experiment were also excluded, as well as those who failed to complete the full survey. Finally, due to a technical error the AS data for 238 participants was corrupted, and these responses were therefore excluded from the analysis. This left us with 2666 full 'valid' responses out of 3024.

Materials and Sales Data

All the attributes and the products used in the experiment were derived from the supermarket scanner data. The scanner data available for this study was provided by one of the four largest U.K. supermarket chains. This data consists of weekly number of products sold and total sale revenue per week per store for fresh milk, tinned tuna and eggs, for a total of 52 weeks and 611 stores. Since the only identifying component of a product was its full name, as displayed on the shelf, the various attributes collated in Table 21 (Appendix 1) were derived from that name. Hence a product labelled "A2 SEMI-SKIMMED MILK 1 LITRE" was coded as a milk product with brand "A2", type "Semi-skimmed", size "1 Litre" and organic "No" (see Figure 3). The price was calculated by dividing the sale revenue by the number of products sold. In order to reduce the correlation between product size and price, price was recoded to "price per unit". For milk products this meant price per pint, price per egg for eggs, and price per 100 grams for tuna. Working from such general data made it impossible to take into account other important attributes such as product placement, special offers and packaging design.

Store id	Week no	Prod cat	Description	Sales	Units sold
001	01	Milk	A2 semi-skimmed milk 1 Litre	405.09	417



Store id	Week no	Prod cat	Brand	Type	Size	Organic	Price/pint
001	01	Milk	A2	Semi-skimmed	1.759 pt	N	£0.55

Figure 3: Example of the sales data including the derivation of the attributes and levels. The milk product's attributes are derived from the description, while the price is given by the sales / units sold and normalised per pint. In the discrete choice experiment, the price displayed is the non-normalised one, in line with what is most salient in supermarkets displays.

Experimental Method

The experiment comprised of three main tasks: a discrete choice experiment (DCE), followed by two direct elicitation methods in random order (constant sum scales and anchored scales). The DCE was always first as this matched most closely the real world choice environment which I was mirroring: people do not usually think deeply about their preferences before making grocery choices. Additionally, there is evidence that warmup tasks which require participants to think about their preferences before making a choice changes the effectiveness of choice experiments (J. Huber et al., 1993), which I wanted to keep constant across participants. Each task targeted a different product category and was randomised across participants: Fresh milk, eggs and tinned tuna. This minimised cross-task contamination. Demographic and attitudinal data was gathered at the end of the session. Following data collection, the analysis involved comparing the attribute importance derived from the various methods and the predictive power of the resulting aggregate models both on a holdout set of choices made by participants, and on samples of the sales data (see Figure 4).

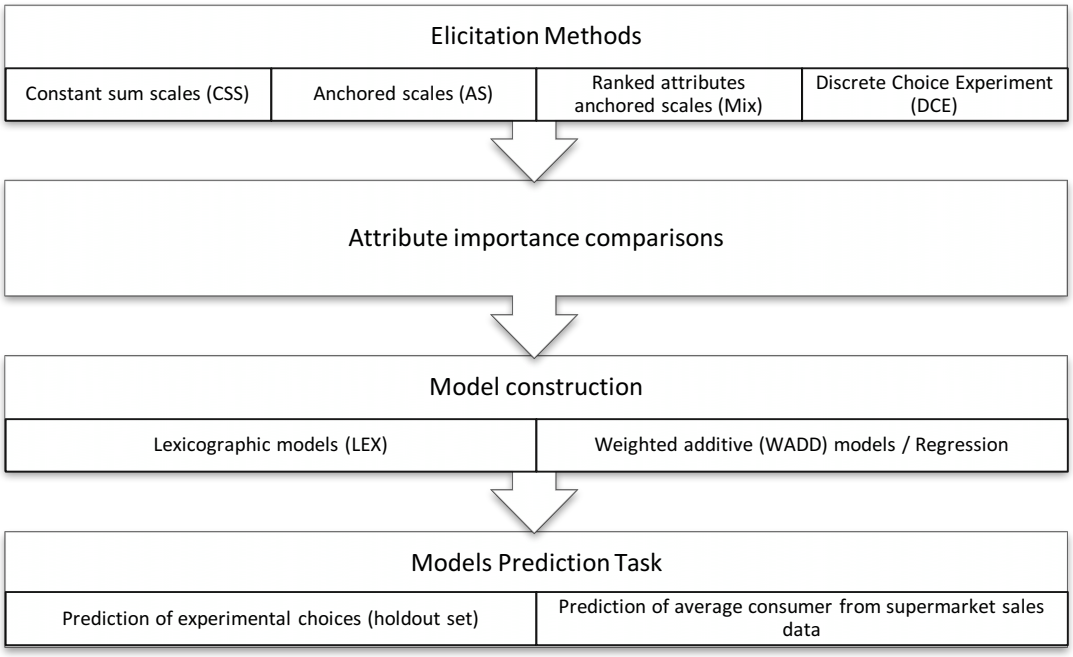


Figure 4: Analysis plan of the experiment. First, all four elicitation methods are compared with regards to their attribute importance measures. Then, each elicitation method is used to to parameterize a lexicographic and a weighted additive model (if possible), which are then compared on two datasets.

Discrete Choice Task

Participants were first presented with a set of three discrete choice tasks for one product category. Each task consisted of three forced choices: most preferred product, second most preferred product and least preferred product; only the first of which was used in the analysis. The instructions invited participants to imagine that they are engaging in their regular grocery shopping and that the products displayed are the ones available at their regular supermarket. The products were displayed on a selectable grid, and consisted of a picture of the product and the full product name (see appendix 2 for an example), in similar style to an online grocery store display. Instead of providing a no-choice option, following each choice respondents stated if they would indeed purchase the selected item. When choosing the second most preferred item, the range displayed was identical to the first apart from the most preferred product, which was labelled as "out of stock". Products shown to the participants during the three choice tasks were derived from the sales data. Two of the ranges of products shown to respondents were directly taken from often

occurring ranges in the data as they were deemed to be more representative of what customers saw when grocery shopping, while the third was formed as a random selection of 14 products. Out of the two representative ranges that participants saw, one had a large number of products (36 items for milk and tuna and 24 items for eggs) and had a more modest number (14 for milk, 18 for eggs, or 9 for tuna). This contrast was used to directly compare distribution of choices in the DCE with the distribution of sales. Prices were dynamically sampled, for each respondent, from the historical distribution of prices of each product with outliers and incorrect data removed.

Direct elicitation tasks

After the choice task, respondents were presented in random order with either the anchored scale (AS) or constant sum scales (CSS) preference measurement for all features and feature levels. Both methods generally followed Huber's (1974) client explicated parameter estimates methodology. The AS task required identifying the most and least important features, or feature levels, assigning to them the values 0 and 100, and then rating the remaining features relative to this scale. For the features of the products (e.g. brand, price, …) participants ranked all the features in preference order, while for the specific feature levels participants were asked to select only the most and least important feature. For example, when asking about fat content of milk, respondents stated which of the five available feature levels (skimmed, semi-skimmed, Jersey, 1% and whole milk) were their most and least important. The next screen displayed all the feature levels again with the participants' selection at the top with a fixed score of 0 for the least preferred level and 100 for the most preferred one. Participants had to rate the other levels, according to their preference, on a 0 to 100 scale using a horizontal slider. The constant sum scale task required participants to distribute 100 points amongst the features and feature levels in line with how important they perceive them to be. In this way, the two direct elicitation methods provided us with specific self-reported weights for each feature and each of its levels. In addition to the features in Table 21, respondents were asked about more general features not used in either the stated preference model or the revealed preference one, namely: 'best quality', 'healthy' and 'locally sourced'. See appendix 2 for specific elicitation question examples.

Model construction

The models derived from the preference elicitation methods were not individual choice models. As for each respondent there were only three 'most preferred' choices and the product groups in all three elicitation methods were for different products, it was not possible to predict individual preferences. Instead, all the answers were averaged and the models were derived for an average consumer. Only the features and levels in Table 21 were part of the models. As shown in Figure 4, three WADD models were constructed from the results of the questionnaires and the choice task. No WADD model was derived for the ranked attributes anchored scales method, because this was reserved for a special purpose LEX model. Direct attribute importance weights from CSS and AS, were used to parameterise two multi-attribute utility models (G. P. Huber, 1974; Srinivasan, 1988). The utility of each attribute (also called part-worth) was calculated as the product of the elicited attribute importance weight and the elicited attribute-level importance. The utility of each specific product was then the sum of its part-worths. The indirect WADD model was obtained by running a multinomial-logistic regression on the DCE results using the mlogit package (Croissant, 2012) in the R programming language. Three LEX models were constructed from the direct attribute importance measures. One was derived from the CSS measures, one from the AS measures and the last one used the direct rankings for the attribute order and the AS measures for the attribute levels, as the AS implied attribute ranking and the actual attribute importance ranking were slightly different. The only continuous variable that the LEX models had was price. To accommodate some flexibility in the perception of price, and to help account for the relatively noisy sales price data, I used a price sensitivity measure. This was set to accommodate indifference between similarly small amounts of money. The parameter (s) is a percentage of the price range of the set of products. Thus when looking at the price attribute, the model does not differentiate between the lowest priced product (costing p_{min}) and the products costing less than $p_{min}+s*(p_{max} - p_{min})$ where p_{max} is the most expensive product available.

Model Predictions

For each of the models, I compared the models' predictions on a holdout sample consisting of the unused half the DCE data and a random sample of 2000 choices, weighted by total weekly sales, from the sales data. In order to make sure

that the results are robust the procedure was conducted 30 times with different, randomly chosen, holdout samples. Each run resampled 50% of the DCE data to parameterise the indirect WADD model while the other 50% was left as the holdout set. Similarly, at each run the sales data used for prediction was resampled.
For each product, I used three different range sizes (except milk, which had only two). In order to combine them into one dataset I added an 'available' feature and padded the shorter ranges with random products having the available feature set to 'N'. A similar manipulation was performed on the sampled sales data in order to have ranges of the same size, per product category, to run the models on. All models chose only amongst the products having the available feature set to 'Y' in each range, except the regression model that assigned a large coefficient to the available feature, in effect making non-available products unchooseable. The hit-rates was used as the main metric in the comparison of predictive validity. This is the number of times that the model predicted the correct chosen product, reported as a percentage of the total number of ranges tested.

Results

Comparison between hypothetical decisions and decisions in the field

Before analysing the attribute importance and the predictive models, I checked for external validity of the experiment by comparing respondents' hypothetical choices to the ones generally present in the field, using the supermarket dataset. Two of the three DCE product ranges were fixed for all respondents, and mirrored exactly two common product ranges from the supermarket scanner data, allowing for a direct comparison. For each product group, I therefore compared directly the DCE market shares for these two ranges to the market shares from the scanner data. Table 2 summarizes the correlation for each static product range, taking into account all the choices made during the DCE or only those where respondents stated their intention to buy the chosen product. Only the Tuna category has a low correlation between the sales data and the choices in the laboratory. Looking more closely at the short range of tuna, the sales data is very polarised: a single product has 47% of the market-share, which is not reflected in the choice data. This might be partly responsible for the low correlation. All the other ranges have moderate to high correlation, which suggest that the decisions taken in the lab are generally similar to the ones taken by consumers in a grocery shopping context.

The difference between the correlations taking into account all the participants' choices, and taking into account only the choices in which they explicitly stated that they would buy the product is minimal. Hence for the rest of the paper, we can take into account all choices of the DCE, not only those which have been labelled with a positive intention to buy.

Buying Intention	Milk		Eggs		Tuna	
	Long	Short	Long	Short	Long	Short
All	0.95	0.76	0.66	0.82	0.55	0.32
Y	0.95	0.77	0.65	0.82	0.54	0.30

Table 2 Correlation between market-shares of the sales data and choices in the DCE for each of the identical ranges (two per product group: a shorter range and a longer one).

Attribute importance

Attribute importance measures were compared between the two direct preference elicitation methods (CSS and AS) and with the importance values derived from the DCE. Following Louviere and Islam (2008), the partial log likelihood associate with each attribute was used as an estimation of importance weight for the DCE.

Consistent with Louviere and Islam's (2008) comparison, correlation between the direct elicitation methods was high (Spearman's $\rho > 0.80$ for all three products), while the correlation between direct and indirect methods was low to moderate (Spearman's $\rho < 0.49$). The pattern was similar when comparing attribute level ordering across methods. In general the lowest correlation between direct and indirect methods was for the quantity attribute (e.g. for CSS: $\rho_{milk} = 0.79$ $\rho_{egg} = 0.18$, $\rho_{tuna} = 0.07$).

Not all elicitation methods resulted in statistically significant differences among the most important attributes, and some even resulted in a different ordering. These differences might not have a high impact on weighted additive models which rely on values, but are very important for lexicographic models which are only concerned with rank, as the results of the hit-rate analysis will show. For LEX models, a small change in an attribute value can change its rank, and therefore change the models' results significantly. This instability in attribute importance ordering across elicitation methods could be indicative of preference heterogeneity in

the population or lack of introspective access by respondents as to the most important attributes that are guiding their choices.

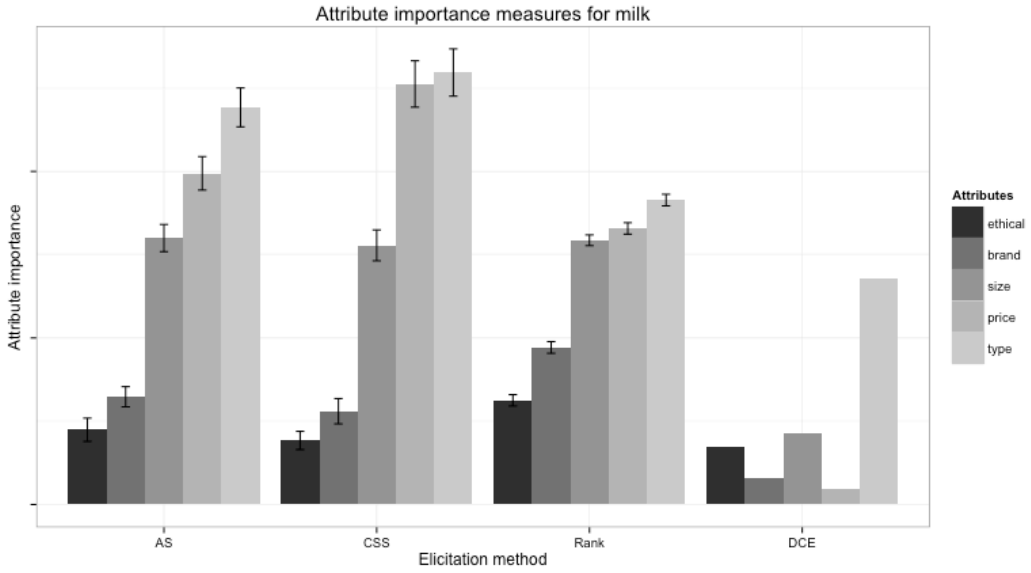


Figure 5: Attribute importance measures for milk products. The three direct preference elicitation methods agree with each other, even though the two most important attributes in the CSS method are not significantly different. The attribute importance derived from the DCE is very different

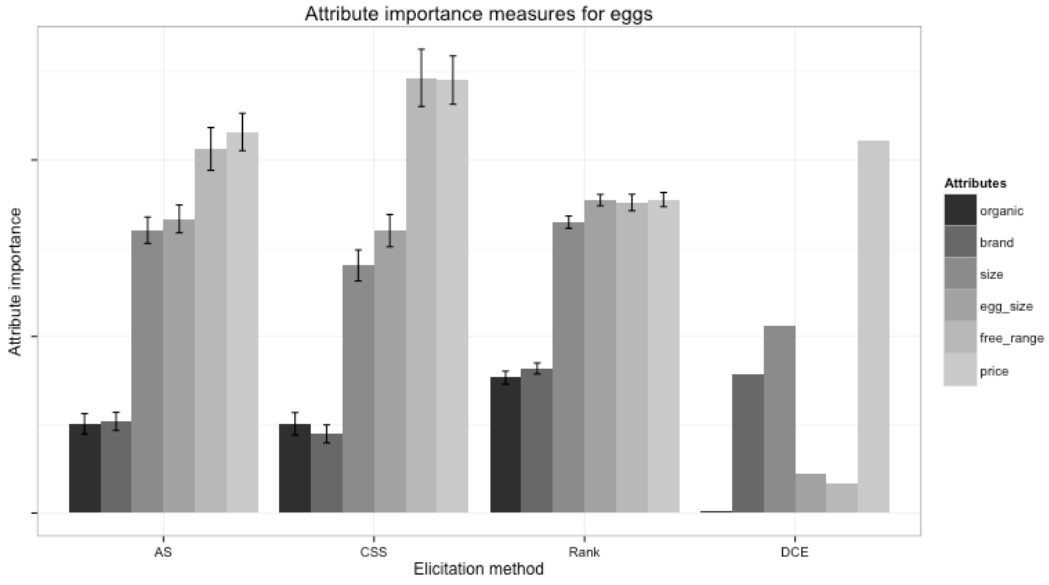


Figure 6: Attribute importance measures for eggs. The pattern is similar to the milk attribute importance, with the agreement between direct measures much higher than with the indirect, DCE, measure. There are many more non-significant differences between consecutive attributes, meaning that the ordering is less robust.

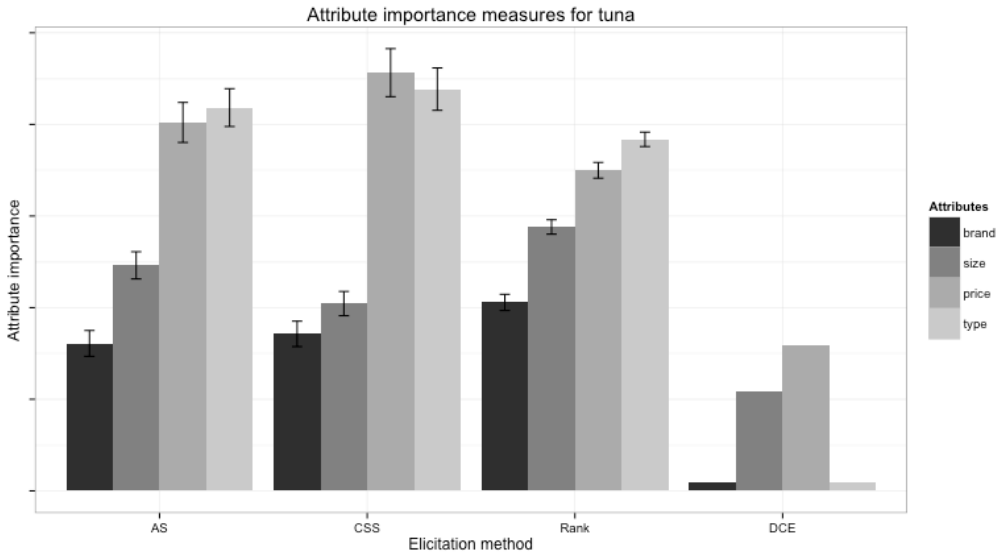


Figure 7: Tuna attribute importance measures. The pattern is quite similar to the eggs, with slightly different attribute orderings provided by both AS and CSS measures, with non-significant differences in importance for the two most highly ranked attributes.

Lex models parameterization

Four distinct LEX models were parameterised from the attribute importance data. One was fully derived from the CSS measures, one from the AS and the third combined the attribute ranking method with the AS feature level ranks. The final model used the attribute importance ordering derived from the revealed preference elicitation method.

In addition to the attribute ordering, another important parameter of the LEX model is the price sensitivity, which sets the level at which prices are perceived as quasi-equal and do not weigh differently on the decision. The price sensitivity was derived by piloting a number of simulations on the data, with the parameter s increasing from 0.005 to 0.45, in steps of 0.05. The optimal setting, yielding the highest aggregated hit-rate for the LEX models was 0.10.

Hit-rate comparisons

Figure 8 summarizes the hit-rates of the models on the DCE holdout set and the resampled scanner data for the three product groups, based on 30 runs. In general, the WADD models elicited directly (denoted AS and CSS) are the worst performers, with prediction rates in some cases lower than the random benchmark. The difference in hit-rates are much more pronounced when validating on sales data.

For example, the models for milk, both in the lexicographic framework and the regression, show hit-rates of around 59%. In the choice task holdout set, this value falls to around 24%. The complexity of the prediction task is exemplified by the low hit-rate of the benchmark random model. Indeed, in each choice task there are up to 36 products that the models choose from. Thus, the hit-rates of 59% and 24% reported above are equivalent to, respectively, an increase of around 19 times and 4 times over the random benchmark.

The hit-rate results on the sales data seem to suggest that the LEX model parameterized using CSS is the most predictive of choices. Two sample paired t-tests between the CSS.Lex and regression hit-rates for milk ($t(29) = -1.2805$, $p > .1$) and egg ($t(29) = -14.7036$, $p < .005$) show no significant differences for the former, but confirm the superiority of the CSS.Lex model for the latter. As the LEX model is very simple, it is surprising that it equates and even dominates the regression model in capturing the average consumer's choices. Given the stark differences between the different LEX models, the elicitation method holds a central role in their effectiveness.

The aggregate results in Figure 8 seem to suggest that the CSS LEX model predicts better than the regression model. Looking at the hit rates per product group we can see that this effect is carried solely by the tuna products. This can be explained by noticing that the regression model is parameterized on the DCE results, and we have seen that the tuna DCE task was the least comparable with the real world choices (see Table 2).

The results of the revealed Lex model are very variable across products. For tuna products, the hit-rate varies between 14.3% and 45.5% with a mean of 27.68%. Although lower than the CSS Lex model, this is higher than the regression, from which it takes the attribute importance orderings. The model's results for milk and eggs are both very close to the random benchmark. This indicated that the attribute importance measured by the direct and indirect preferences are not interchangable and therefore might be different in nature. While the indirect elicitation methods elicited through the DCE seem to be well suited for WADD models, the direct elicitation methods fare much better with the simpler LEX models.

	Milk		Eggs		Tuna	
	DCE	DB	DCE	DB	DCE	DB
Regression	**24**	**59**	**22**	**42**	**17**	13
	(0.15)	*(0.36)*	*(0.12)*	*(0.25)*	*(0.15)*	*(0.29)*
Anchored scales	14	1	3	0	12	3
	(0.13)	*(0.04)*	*(0.08)*	*(0.02)*	*(-0.14)*	*(0.07)*
Constant sum scaling	16	18	4	0	12	3
	(0.13)	*(0.18)*	*(0.06)*	*(0.03)*	*(0.13)*	*(0.07)*
Random	6	3	5	5	7	3
	(0.10)	*(0.06)*	*(0.11)*	*(0.11)*	*(0.14)*	*(0.07)*
Rev.Lex	12	1	12	6	11	28
	(0.53)	*(0.06)*	*(0.15)*	*(0.88)*	*(0.09)*	*(2.61)*
Mix.Lex	23	**60**	13	1	15	9
	(0.16)	*(0.24)*	*(0.11)*	*(0.04)*	*(0.13)*	*(0.11)*
AS.Lex	23	**60**	13	1	15	9
	(0.16)	*(0.24)*	*(0.11)*	*(0.05)*	*(0.13)*	*(0.11)*
CSS.Lex	24	**60**	21	**44**	15	**42**
	(0.16)	*(0.24)*	*(0.14)*	*(0.18)*	*(-0.13)*	*(0.24)*

Table 3: Hit rate results for model predictions on a holdout set of experimental choices (DCE) and on the supermarekt scanner data set (DB). Standard errors are shown in brackets while the highest hit rates are in bold, for ease of identification. On the DCE prediction, regression outperforms all the other methods. On the DB prediction, the results are quite close to the best LEX model, except for the tuna category. Results for the LEX model are very dependent on the elicitation method used.

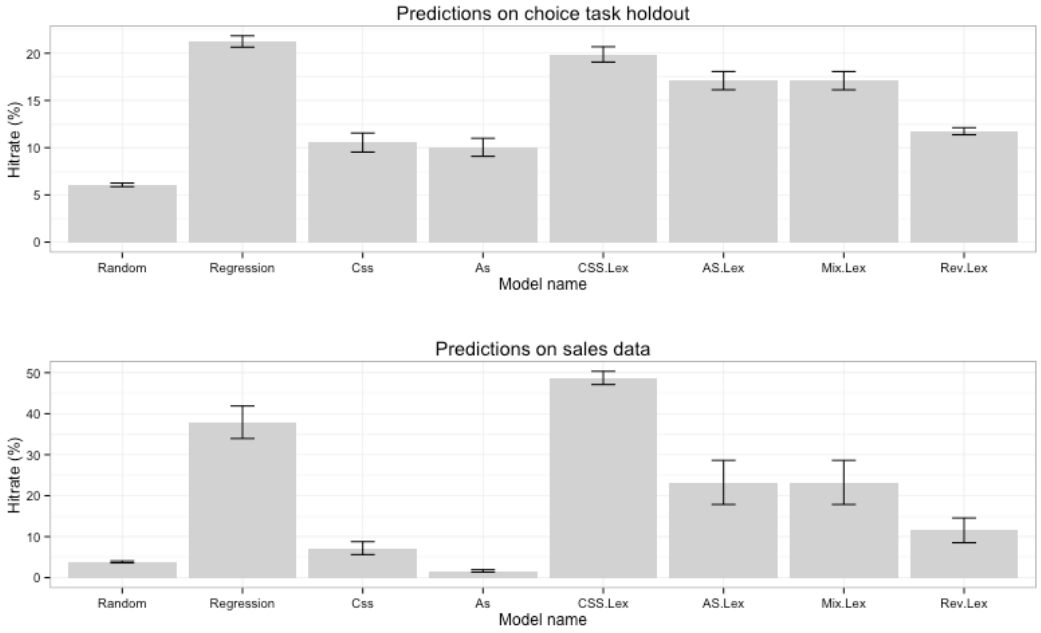


Figure 8 Summary of the predictive models' hit-rate on samples of the DCE data and the supermarket scanner data (30 runs). Although the regression model on the sales data looks lower than the CSS LEX model, this is only due to its low predictive validity on the tuna data. The raw data in Table 3 shows that in general regression yields very close results to the best LEX model.

Stability of the models

Since the attribute importance ordering was not always statistically different for each elicitation method, I tested the stability of the models by parameterizing all the models on only 50% of the participants on each run. Thus, I re-ran all the models 30 times, parameterizing them on random subsamples of half the respondents. The expectation was that, since direct preference elicitation is not robust, it would result in much less stable hit-rate measures, especially for the LEX models, expressed through greater standard errors. Conversely, if the preference elicitation is robust, we would expect the results of the LEX models to be very similar in both hit rate and variation to the ones in Figure 8. The results of match our expectations with one interesting exception. The results for the milk category were virtually unchanged (see for Figure 19 details), and the errors bars were just as small as for the models based on the full population. This is an indication that preferences for milk are quite homogeneous in the population, and that there is a certain level of introspective access to these preferences, since the hit rate is high. The on average lower hit rates for LEX models show that although still much better than the direct elicitation

WADD models, the DCE and regression are still the most reliable preference elicitation method in terms of choice prediction.

Discussion and conclusions

The main aim of this experiment was to explore the limits of introspection, specifically for commonly purchased products. In doing so, I examined the consistency of the importance ranking of attributes for such products using three direct preference elicitation methods. I also compared the accuracy of the direct elicited methods and an indirect elicitation method (a discrete choice experiment) in predicting choices, using two different modelling frameworks. Before comparing the methods amongst themselves, I made sure that the choices made experimentally mirrored the real world behaviour, by comparing experimental market-shares with ones extracted from a large supermarket scanner data set. Following this validation, the different elicitation methods were compared on their resulting attribute importance scores. Finally, these scores were used to parameterize two types of predictive models which were compared in terms of predictive ability.

The comparison of choices made by participants and the market confirmed a good match between choices of milk and eggs in the experiment and in the supermarket. On the other hand, the correspondence of tuna choices was significantly lower. This could be due to the increased reliance of consumers on special offers, visual cues or product shelving in the supermarket. Due to the inherent simplification of the DCE and the aggregate nature of the supermarket data, these effects were not able to be captured in the experiment. Thus, the hypothesis of a limited impact of hypothetical bias in discrete choice experiments (Adamowicz et al., 1998) was confirmed for milk and eggs choices.

In comparing the attribute ordering of the elicitation methods, I observed a high correlation amongst direct methods as reported elsewhere in the literature (Louviere & Islam, 2008). More importantly, in the lexicographic framework these small changes resulted in much larger changes in predictive validity. Due to the nature of LEX models, small changes can have very big consequences, by affecting the model's always decreasing consideration set at each step. For example, the AS method of parameterizing the egg model lead to a hit-rate close to 0% on the sales data, while the similar CSS model predicted more than 40% of these choices,

because the most important attribute is price in the former model and free-range in the latter. Since the various elicitation methods resulted in different attribute ordering, reliable elicitation of the optimal lexicographic ordering was not possible with the direct elicitation methods. Additionally, the direct elicitation methods were not able to always differentiate the most important attributes in a statistically significant manner. This could be indicative of a high level of preference heterogeneity in the population. Yet the high hit-rates of some of the lexicographical models indicate the existence of a lexicographic pattern to the average consumers' purchases. Hence it seems that there are clear limitations in respondents' introspective access, since the direct elicitation methods are not able to consistently extract the preference ordering required for these optimal LEX models. Another indication that preferences, even for familiar and experienced products, are not easily accessible is shown by the instability of the lexicographic models when parameterized from a subgroup of respondents. When repetitively sampling large subgroups of respondents, the attribute importance ordering and derived LEX models varied much more in terms of predictive ability, unlike the regression model. This shows an instability in the directly elicited preferences of the average consumer, and, along with the variability of attribute orderings elicited from the respondents, points towards a lack of introspective access. The milk category seems to be an exception to this, since although the attribute importance ordering was not always statistically significant, preferences were stable enough that all direct elicitation methods, even for subgroups, provided a highly predictive LEX model. This could be due to the increased familiarity of this product as milk was the most frequently bought category, and there is therefore the possibility that for simple very habitual purchases, preferences are indeed accessible. Another possibility is that given the familiarity of the choice scenario, participants have access to their previous choices and are therefore remembering their past behaviour as opposed to accessing their preferences.

Taking the psychologically plausible view that absolute values are not stored in the brain (Vlaev et al., 2011), lexicographic models which do not require absolute values should be easier to elicit. When paired with a predictive model, which also does not rely on absolute values, we should have a better prediction of choices than the weighted additive models. This is indeed the case in the experimental data: the

WADD models based on direct elicitation techniques predict choice very poorly compared to the LEX ones. More specifically, the CSS LEX model predicted choices on the sales data as well as, or better than, regression model. This reinforces the plausibility that that preferences do not necessarily require values to be stored in the brain which is important for both choice prediction and preference elicitation. Unfortunately, since even the most predictive of the lexicographic models did not have a statistically significant attribute ordering for all products, none of the elicitation methods are robust, which indicates clear limitations in the direct preference elicitation methodology, and the introspective access that the respondents have to their preferences.

The indirect elicitation technique and regression model, which is the most commonly used predictive model, generally performed slightly better than the best lexicographical model, except for in the category of tinned tuna. As the regression model uses choices in the experiment to generalize to the behaviour of the average consumer to the sales data, this can be explained by the low correspondence between tuna choices made in the two contexts. Interestingly, unlike the mixed support for the direct elicitation method using weighted additive models (summarized in Table 1), this study clearly points to the inferiority of this approach in the context of habitual decisions. Conversely, in a lexicographic framework, directly elicited preferences do have some predictive advantage over ones derived from a DCE. In addition to taking into account the limited cognitive capacity that people employ for habitual decisions, these lexicographical models and their directly elicited preferences are also simpler to communicate to non-experts.

Attribute importance measures can be easily derived from a regression model using Louviere and Islam's (2008) method. Yet, the lexicographical model derived from these indirectly elicited attribute importance has a very low hit-rate in general. This finding raises questions concerning the interpretability of attribute importance measures from both approaches. In the lexicographic framework, it is possible to say that the most important attribute is indeed the one that has the most effect on the final choice outcome as it is used in a non-compensatory way to narrow the consideration set. On the other hand, the indirectly elicited attribute importance ordering does not translate well into this framework. Indeed, the attribute importance for milk derived from the regression, which itself is very predictive of choices, leads

to an unpredictive lexicographical model. The notion of attribute importance is not equivalent in both frameworks, and therefore interpreting participants' answers to preference elicitation questions is not straightforward. Marketers and researchers should therefore use caution when comparing directly and indirectly elicited preferences.

The low predictive validity of the directly elicited lexicographic model does not in itself mean that the proposed approach to elicit preferences for habitual decisions is limited by introspective access. Indeed, although deriving attribute importance from the regression and using them in a LEX model does not perform well, a different approach exist for deriving lexicographical models from an indirect preference elicitation method (Kohli & Jedidi, 2007; Yee et al., 2007). Dieckmann et al.'s (2009) comparison between this approach and conjoint approach has shown an advantage for the latter, but since the decision context was ski-jackets, which are not classified as habitual decisions, this could be researched further.

In conclusion, using an experimental approach coupled with a large supermarket sales dataset, I set up to investigate people's introspective access to their preferences by comparing direct and indirect preference elicitation methods. I compared the validity of these methods by predicting aggregate choices using two modelling frameworks (weighted additive models and lexicographical ones). The results were somewhat mixed, with people's access to their preferences for milk products quite high, unlike their preferences for the other two frequently bought products: eggs and tinned tuna. The comparison also highlighted the limitations of the direct elicitation approaches with weighted additive models in this context, showing that lexicographical models of habitually bought products have generally a much higher predictive validity. The commonly used discrete choice experiment and regression model were still the most robust and consistent predictors of choice, justifying their wide use. As the validation of the models was carried out in an aggregate manner, by predicting the choices of an average consumer, further research should validate these findings at an individual level. In addition to exploring more fully people's limits in introspecting their preferences, this will have the added benefit of assessing the applicability of the simpler and less time consuming lexicographic preference elicitation methodology to individual decisions.

Chapter 4. Exploring the preferences and choices of ethical consumers for frequently purchased grocery products in the U.K.

Abstract

The exploration of consumer preferences is often conducted through direct elicitation methods, be they self-reports or interviews. Concerns with the misalignment of directly elicited preferences with actual behaviour has encouraged the use of indirect methods such as conjoint analysis or discrete choice experiments. It might be expected that such misalignment may be especially evident regarding 'ethical' purchasing considerations, which people might be expected to over-report. In the present study, I examine the consistency of direct preferences for ethical, occasionally ethical and non-ethical consumers, for two frequently purchased product categories: fresh milk and eggs. I also explore the degree to which direct and indirect preferences converge for these three consumer groups. All consumers, including those not identifying as ethical, exhibited a gap between their directly and indirectly elicited preferences. Directly elicited preferences of ethical consumers were less homogeneous than for other consumers, indicating a diverse interpretation of ethicality for those consumers. Additionally, the similarity of the indirectly elicited preferences for all consumers suggests the limited impact that general ethical concerns have on specific choices for habitual purchases.

Introduction and Literature review

Ethical consumers are usually defined as consumers who augment their decision making strategies with ethical concerns (Harrison, Newholm, & Shaw, 2005) in addition to taking account of their preferences regarding intrinsic properties of the product. Thus, in addition to taking into account traditional product attributes such as price and quality, these consumers consider ethical attributes such as 'fair-trade' and 'organic' when making a purchase. These consumers are also defined as having a more deliberative decision making context (Crane & Matten, 2005). The

classification of consumers as ‘ethical’ or ‘non-ethical’ often requires self-reporting, as behavioural cues are inexact. For example, a consumer buying organic products may do so for ethical reasons, such as not wanting to harm the environment with pesticides, and can therefore be classified as an ethical consumer. Yet, many organic consumers cite the non-ethical factors of taste and health, as driving their preferences (D. Shaw, Grehan, Shiu, Hassan, & Thomson, 2005; Zanoli & Naspetti, 2002). The aim of this study is to provide a better understanding of ethical consumers by looking at the consistency of their directly elicited preferences as well as the degree to which these align with their indirectly elicited ones, specifically in the context of habitual purchases such as milk and eggs. The paper is organized as follows. First, some general background about direct and indirect preference elicitation methods, as well as ethical consumers will be presented. I then describe the multi-part direct and indirect preference elicitation experiment, before presenting the results and, finally, drawing general implications of the findings.

The self-reporting of preferences and behaviour, through direct elicitation methods, has well-known limitations. It has often been argued that people do not have as much introspective access to the working of their own mind, as they consciously think they do (Carruthers, 2009). For example, Johansson, Hall and Sikström (2008) have shown that when asked about reasons behind some of their decisions, people can give post-hoc explanations which cannot be true. In their study, respondents were to make a choice between which of two pictures they prefer. They were then deceived using sleight-of-hand and presented with the picture they did *not* choose while being asked to explain their choice. The majority of respondents did not notice that the picture was not the chosen one, and justified their choice using elements present in the displayed image. The decreased confidence in the validity of people’s introspective process (Nisbett & Wilson, 1977a) has led to the proliferation of indirect preference elicitation methods (Green & Rao, 1971; Green & Srinivasan, 1978). Moreover, even self-reports of one’s own habitual behaviour have been shown to be unstable (Schwarz & Oyserman, 2001). For example, when reporting the frequency of a behaviour, the response scale shapes the answers. Respondents have reported watching more television (Schwarz et al., 1985), suffering medical symptoms more frequently (Schwarz & Scheuring, 1992), or buying more soft drinks (Menon, Raghubir, & Schwarz, 1995) when using a scale with more high frequency items. Regardless of these limitations, direct elicitation

methods are still widely used by marketers and academics, partly because there are relevant straightforward to measure, when compared with more indirect preference elicitation methods, which aim to reveal preferences indirectly through people's choices.

Direct elicitation methods encompass interviews as well as various survey techniques which require consumers to rank or rate different attributes in terms of how much weight they hold in their decision making process (e.g. Chrzan & Golovashkina, 2006). For example, when asked about her milk purchasing, a consumer might state that price has an importance score of 40, carton size 15, and fat content 20. This means that price is the most important factor in the decision, followed by fat content and finally size. Conversely, revealed preference techniques rely on inferring the importance of the attributes by analysing people's choices between, or ratings of, products. One of the most common example of these techniques is the discrete choice experiment (DCE; Wittink et al., 1994). In a DCE, consumers are faced with a range of products and have to choose which one they prefer, or which one they would purchase. After a number of choices have been made from different ranges, a multinomial logistic regression can be calculated and the weight of the different attributes extracted.

When using direct elicitation methods, the underlying assumption is that consumers have sufficient introspective access to report their preferences and actions, i.e. that what consumers say is not different from what they subsequently do. Unfortunately, this assumption does not always hold, leading to an attitude-behaviour gap (Boulstridge & Carrigan, 2000; Sheeran, 2002). For example, in the context of ethical purchasing, Auger and Devinney (2007) have investigated this gap by comparing direct survey elicitation with a DCE. They concluded that survey questions overstate ethical concerns and are therefore not a reliable source of purchase intention data. More generally in preference elicitation studies, consumers have been known to choose products whose attributes they previously rated as unacceptable (Green et al., 1988). Consumers are also prone to misreport what influences their decisions. This is exemplified by Muller, Lockshin and Louviere's (2009) study in which respondents rated as relatively unimportant visual cues such as packaging, when asked about their wine preference using a direct elicitation method. An analysis of the same respondents' choices in a DCE revealed that these visual attributes had a significantly more important effect on choice outcome. In the

context of ethical purchasing, the misreporting of attitudes can also be due to the social desirability bias (DeMaio, 1984). This occurs when respondents change their answer to questions in order to portray themselves in better light. Indirect preference elicitation methods also suffer from people's introspective limitations, as measured preferences can be tangled up with external influences. An item's position in a range, for example, influences its appeal (Bar-Hillel, 2015) yet people do not report knowledge of this effect (Wilson & Nisbett, 1978). These position effects are also biasing for DCEs and can lower their validity. When analysing choices made experimentally, if the position effects are significant and differ greatly from the real world choice context, then the preferences elicited might not reflect actual behaviour.

Eliciting ethical consumers' preferences can be used to understand what attributes act as barriers to consumption and what attributes are most positively viewed by them. These preferences can also be used to help design new products, marketing campaigns and to predict brand market-share. Self-reports, interviews, focus groups and revealed methods have been widely used to explore these consumers' preferences. Product-specific studies (e.g. Becker, Tavor, Friedler, & Bar (Kutiel), 2016; Johnston, Wessells, Donath, & Asche, 2001; Mesías, Martínez-Carrasco, Martínez, & Gaspar, 2011), as well as more general studies reveal some similar trends. For example, price, availability and convenience are often cited as barriers to ethical consumption while attributes which are associated with ethical consumption include good taste, better quality, more healthy and concern for the environment (Essoussi & Zahaf, 2009; Joshi & Rahman, 2015; A. K. Moser, 2016; Padel & Foster, 2005; Sirieix, Kledal, & Sulitang, 2011; Zepeda & Deal, 2009). Hence ethical consumers are often differentiated by their lower price sensitivity and higher willingness to pay for ethical products. Additionally, given these consumers' deliberative approach to making purchasing decisions (Crane & Matten, 2005), we might expect that these directly elicited preferences should be more closely reflected in their choices. Yet this effect is much smaller in practice than self-reports suggest (Carrington, Neville, & Whitwell, 2010). Demographic variables are generally less useful in differentiating ethical from non-ethical consumers (Roberts, 1995). Revealed preference techniques have also been used in this context. For example, De Pelsmacker, Driesen and Rayp (2005) clustered respondents to a willingness-to-pay

survey for ethical coffee and showed very little demographic difference between the average consumer and the ethical ones.

In summary, this study aims to use both revealed and stated methods to characterize ethical consumers, their preferences and their choice behaviour in a realistic shopping context. The consistency of direct preferences amongst the three groups will inform us on their cohesiveness, while the comparison between directly and indirectly elicited preferences will uncover if any of the groups has more introspective access to what drives their behaviour.

Methods

The Participants

Participants were recruited on an online commercial panel provider (Belindi) and were U.K. residents screened to be fluent English speakers older than 18 years old who frequently buy milk, tinned tuna and eggs. The median age was 52 years and 39% of respondents were male. Performance in the experiment was not incentivized, and participants were urged to give their true preferences in return for points in a loyalty program. Answers that were deemed inconsistent were excluded and not rewarded. Inconsistency was measured by asking participants for their year of birth at the beginning of the survey and their age at the end, answers that differed by more than 2 years were inconsistent. Additionally, participants that took less than 10 minutes to complete the experiment were also excluded. Finally, due to a technical error the Anchored Scales data for 238 participants was corrupted, and these responses were therefore excluded from the analysis. This left us with 2666 full valid responses out of 3024.

Experimental Method

The experiment comprised of four main tasks: a discrete choice experiment (DCE), a free elicitation task, followed by two direct elicitation methods in random order (constant sum scales and anchored scales). The DCE was always first as this matched most closely the real world choice environment which I was mirroring: people do not usually think deeply about their preferences before making grocery choices. Additionally, there is evidence that warmup tasks which require participants to think about their preferences before making a choice changes the effectiveness of choice experiments (J. Huber et al., 1993), which I wanted to keep constant across

participants. Each task targeted a different product category and was randomised across participants: Fresh milk, eggs and tinned tuna, or order to minimise cross-task contamination. Demographic and attitudinal data was gathered at the end of the experimental session.

Discrete Choice Task

A supermarket scanner data set was used in order to add realism to the DCE task by ensuring products and their prices were drawn from real world distributions, and therefore familiar to most consumers. Participants were first presented with a set of three discrete choice tasks for one product category. Each task consisted of three forced choices: most preferred product, second most preferred product and least preferred product; only the first of which was used in the analysis. The instructions invited participants to imagine that they are engaging in their regular grocery shopping and that the products displayed are the ones available at their regular supermarket. The products were displayed on a selectable grid, and consisted of a picture of the product and the full product name, in similar style to an online grocery store display. Instead of providing a no-choice option, following each choice, respondents stated if they would indeed purchase the selected item. When choosing the second most preferred item, the range displayed was identical to the first apart from the most preferred product, which was labelled as "out of stock". Products shown to the participants during the three choice tasks were derived from the sales data. Two of the ranges of products shown to respondents were directly taken from often occurring ranges in the data as they were deemed to be more representative of what customers saw when grocery shopping, while the third was formed as a random selection of fixed length. Out of the two representative ranges that participants saw, one was chosen with a large number of products (36 items for milk and tuna and 24 items for eggs) and one with a more modest number (14 for milk, 18 for eggs, or 9 for tuna). These were used to directly compare distribution of choices in the DCE with the distribution of sales. All randomly formed ranges comprised 14 products. In all cases, prices were dynamically sampled, for each respondent, from the historical distribution of prices of each product with outliers and incorrect data removed.

Free elicitation task

In the free elicitation task, respondents were given 6 empty text boxes and asked to write down the most important attributes affecting their purchase decisions.

The product class used was the same one as in the DCE. Subsequently they were asked to rank the attributes in terms of importance, but since the pattern closely followed the order in which they wrote down the attributes, this data was not used.

Direct elicitation tasks

After the free elicitation task, respondents were presented in random order with either the anchored scale (AS) or constant sum scales (CSS) preference measurement for all features and feature levels. Both methods generally followed Huber's (1974) client explicated parameter estimates methodology. The AS task required identifying the most and least important features, or feature levels, assigning to them the values 0 and 100, and then rating the remaining features relative to this scale. For the features of the products (e.g. brand, price, …) participants ranked all the features in preference order, while for the specific feature levels participants were asked to select only the most and least important feature. For example, when asking about fat content of milk, respondents stated which of the five available feature levels (skimmed, semi-skimmed, Jersey, 1% and whole) were their most and least important. The next screen displayed all the feature levels again with the participants' selection at the top with a fixed score of 0 for the least preferred level and 100 for the most preferred one. Participants had to rate the other levels, according to their preference, on a 0 to 100 scale using a horizontal slider. The Constant Sum Scale task required participants to distribute 100 points amongst the features and feature levels in line with how important they perceive them to be. In this way, the two direct elicitation methods provided us with specific self-reported weights for each feature and each of its levels.

Results

Who are ethical, non-ethical and occasional consumers?

The N = 2666 consumers were divided into three groups depending on their self-assessment of how often they buy ethical products. Respondents who answered 'always' or 'often', were labelled 'ethical consumers' and accounted for 23% of the sample (N=613), and those who reported occasionally purchasing ethical products (57%, N=1517) were labelled 'occasional' ethical consumers. The final group was made up of non-ethical consumers who reported never buying ethical products (20%, N=536). Demographics and attitudes are summarised in Table 4 and Table 5.

	Gender		Age			Higher Education		Full time Employment	
	F	M	18-35	36-55	55+	Yes	No	Yes	No
Ethical	66%	34%	23%	39%	38%	45%	55%	41%	59%
Occasional	61%	39%	15%	40%	46%	35%	65%	36%	64%
Non-Ethical	53%	47%	14%	40%	46%	20%	80%	36%	64%

	Yearly Income				
	< £25,000	£25,000 - £35,000	£35,000 - £50,000	£50,000	Withheld
Ethical	37%	15%	17%	17%	14%
Occasional	41%	19%	14%	11%	15%
Non-Ethical	43%	19%	15%	10%	13%

Table 4: Demographic information for ethical occasional and non-ethical consumers.

	ethical	occasional	non-ethical
Frequency of purchase (1 = never, 4 = always)			
Organic	2.33	1.69	1.19
Locally produced	2.85	2.34	1.95
Fair trade	2.81	2.15	1.57
I usually buy the cheapest product*	3.31	3.70	4.11
I usually buy the best quality product*	5.09	4.40	4.18
Organic products are usually overpriced	5.26	5.62	5.91
I trust the organic labels on products at the supermarket*	4.60	3.83	3.09
I generally buy the same product every week	4.69	4.62	4.72
I often buy new grocery products I have not tried before*	4.81	4.52	4.26
I am very selective about the food that I buy*	5.36	4.90	4.72
People close to me believe that ethical food:			
Tastes better*	4.91	3.94	3.23
Is healthier*	5.01	4.12	3.30
Is worth paying more for*	5.20	3.93	2.82
I believe that buying ethical food is:			
Unimportant*	2.49	3.47	4.49
Difficult*	3.24	3.80	4.13
Unpleasant*	2.25	2.92	3.70
Bad	2.08	2.75	3.61
Foolish	2.14	2.94	3.92

Table 5: Attitudes expressed by the three different consumer groups, especially with regards to organic and ethical concerns. significant differences amongst groups, with $p < 0.001$, based on a MANOVA. Other variables are either non-normal or were not significant.*

Ethical consumers are, in general, more educated and have a higher income than the other two groups. They also report buying more organic, locally produced and fair trade products than their non-ethical counterparts, with organic concerns seemingly the least important.

In terms of attitudes, the results are unsurprising. Ethical consumers report being less price-sensitive, more trusting of organic labels and live in an environment which puts a greater value on ethical food. Ethical consumers also reported shopping at more prestigious and hence expensive retailers. Interestingly, all three groups reported actively looking for special offers, even the groups who regarded price as less of a barrier for consumption. This general characterisation of ethical consumers, consistent with the existent literature (e.g. Sachdeva, Jordan, & Mazar, 2015; D. Shaw et al., 2005) increase our confidence that the questions were answered in a consistent manner and therefore that the data collected is of high quality.

Comparison between hypothetical decisions and decisions in the field

Before analysing the indirectly elicited preferences, the DCE was checked for external validity by comparing respondents' hypothetical choices to the ones generally present among real-world consumers, using the supermarket dataset. Two of the three DCE product ranges were fixed for all respondents, and mirrored exactly two common product ranges from the supermarket scanner data, allowing for a direct comparison. For each product group, I therefore compared directly the aggregate DCE market shares for these two ranges to the market shares from the scanner data. Table 2 summarizes the correlation for each static product range, taking into account all the choices made during the DCE or only those where respondents stated their intention to buy the chosen product. Only the Tuna category has a low correlation between the sales data and the choices in the laboratory. This is especially true for the short range which, in the sales data is very polarised, with one product having 47% of the market-share. This skewness is not reflected in the choice data and might therefore be responsible for the low correlation. Additionally, tinned tuna had the least clear ethical dimension, as every single product had some form of ethical certification (e.g. "line caught", "dolphin friendly" …), leaving no easily identifiable "unethical" tuna. For these reasons, the tuna results were dropped from the analysis. All the other ranges have moderate to high correlation, which suggest that the decisions taken in the lab are generally similar to the ones taken by consumers in a

grocery shopping context. The difference between the correlations taking into account all the participants' choices, and taking into account only the choices in which they explicitly stated that they would buy the product is minimal. Hence for the rest of the paper, we can take into account all choices of the DCE, not only those which have been labelled with a positive intention to buy.

Buying Intention	Milk		Eggs		Tuna	
	Long	Short	Long	Short	Long	Short
All	0.95	0.76	0.66	0.82	0.55	0.32
Y	0.95	0.77	0.65	0.82	0.54	0.30

Table 6 Correlations between market-shares of the sales data and choices in the DCE for each of the identical ranges (two per product group: a shorter range and a longer one). Due to the low correspondence between hypothetical choices and the aggregate market share, the tuna products were not included in the subsequent analysis.

Directly elicited preference analysis

To compare the different groups in terms of directly elicited preferences, the results of the two direct elicitation methods were aggregated. In order to allow this, the AS data was rescaled to add up to 100 for each respondent. This resulted in one direct preference value for each participant, regardless of the direct elicitation which was used to rate the specific product class. The results for milk indicated that the values obtained from both occasional and ethical consumers were more influenced by the type (i.e. fat content) of milk than non-ethical consumers who rated price as the most important attribute. The ordering of the other attributes is quite similar among the different groups. Surprisingly, health and organic concerns are rated as having low importance for all groups, even ethical consumers. For eggs, free range was rated as most important for ethical consumers, much higher than price, while for non-ethical consumers the trend was inversed. Occasional consumers showed no significant preference for price over free range. Organic concerns were again rated quite low, although significantly higher for ethical consumers than for the two other consumer groups. Finally, for both product groups, brand was not rated as important, showing that even if some brands are marketed as more ethical than others, consumers express disregard for this information.

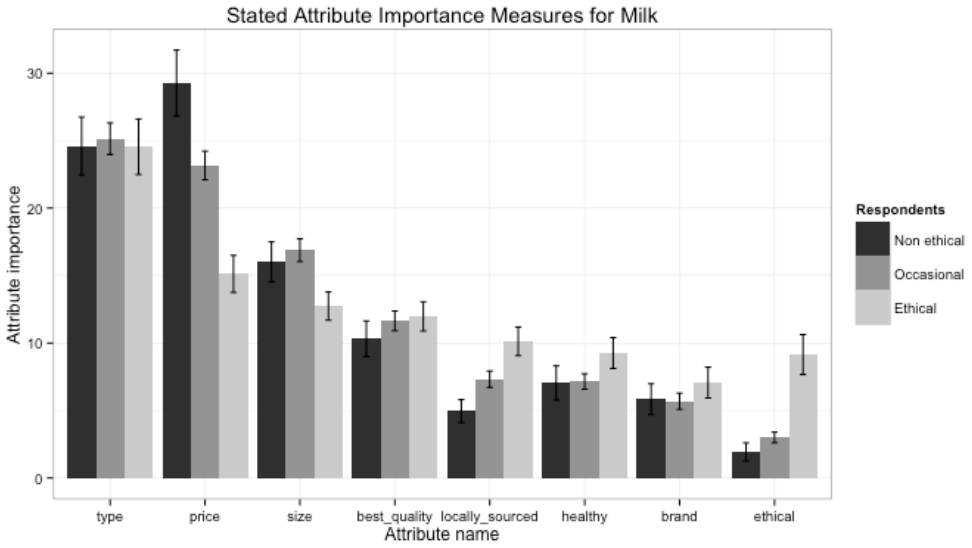


Figure 9: Milk stated attribute importance measures for all respondent groups. The most glaring differences between ethical and non-ethical consumers are seen in the importance of price, locally sourced and organic attributes.

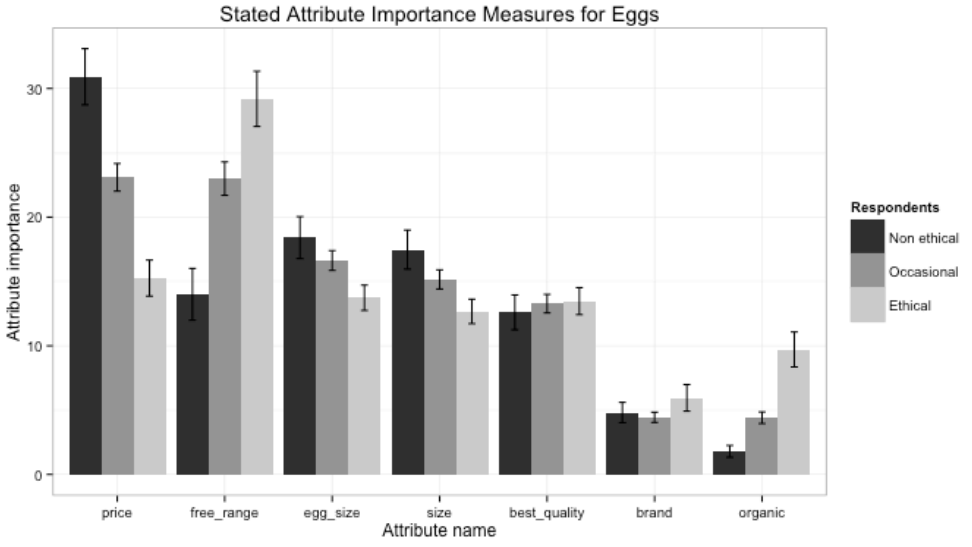


Figure 10: Eggs stated attribute importance measures for all respondent groups.

Measuring within-individual consistency

As part of the anchored scales task, consumers were asked to rank the different attributes in terms of importance, after which, using the most and least important attribute as anchors, they were asked to rate their importance numerically between 0 and 100. This double measure of attribute importance for the same

product category by each individual allows us to check each individual's stated preference consistency. The rank-correlation[4] between both measures, averaged per consumer group is presented in Table 7.

	Ethical	Occasional	Non-Ethical
Milk	0.73	0.80	0.81
Eggs	0.79	0.83	0.81

Table 7: Average correlation between ranking and anchored scales stated preferences, per consumer group.

A one-way ANOVA was conducted to compare the effect of self-reported ethicality on individual's stated preference consistency. Milk displayed the most striking differences ($F(2,863) = 14.67$, $p < .001$), while the effect was smaller for eggs ($F(2,950) = 3.538$, $p = .02$). Ethical consumers were individually the least consistent in their stated preferences. A post hoc Tukey test confirmed that the differences, for milk, were significant between ethical consumers and the other two groups ($p < .001$) while there were non-significant differences between occasionally ethical and non-ethical consumers. For the eggs, the only significant difference was between ethical and occasional consumers ($P = .022$), and even that difference was small.

Measuring preference homogeneity

In order to quantify the differences in stated preference between the groups, correlation values were calculated between every participant between every group. For example, the preferences of every ethical consumer were correlated with the preferences of every occasionally ethical consumer, and so on. The average value of these correlations quantifies the difference in directly elicited preferences among groups. Within-group correlations serve as a measure of preference heterogeneity: the larger the correlation, the more similar the preferences are within the group.

	Ethical	Occasional	Non-ethical
Ethical	**0.162**		
Occasional	0.216	**0.344**	
Non-ethical	0.182	0.335	**0.361**

[4] An alternative analysis was conducted taking into account a different way of ranking the Anchored Scales data, whereby ratings which differed by less than 5 points were considered equal. The results still held, indicating the robustness of the effect.

Table 8: Average correlations of stated preferences between every participant, aggregated by ethical group and averaged across all products[5]. The values in bold represent the within group measure of heterogeneity.

In general, we would have expected the within group correlations to be higher than the between group correlations, yet this pattern is not clear in the results. Non-ethical consumers and occasionally ethical consumers show the greatest preference homogeneity, while the ethical consumers have low within group correlation. Non-ethical consumers are most closely correlated with occasionally ethical consumers. Ethical consumers' preferences, on the other hand, are very heterogeneous. This indicated a lack of consistent shared values within this group. Alternatively, the lack of homogeneity within the directly elicited preferences of ethical consumers could indicate that general ethical values do not translate into ethical attitudes for the specific products under consideration, hence maybe ethical concerns are product-specific.

Free elicitation of Stated Preferences

The participants' free responses were cleaned up and re-coded into a number of attributes (13 for milk and 12 for eggs). A subset of these attributes corresponded to the ones given in the prompted stated elicitation methods, and thus allowed for a direct comparison between the two methods. Since the median number of stated free attributes was 3, and the results are aggregated by consumer group, I used the frequency of each attribute as a measure of its importance.

[5] The same pattern of results is seen when looking at each product separately. The only difference is that for milk ethical consumers' preferences are more similar to those of non-ethical whereas for eggs the reverse is true. In both cases, the preferences of ethical consumers are still most correlated with those of occasional consumers.

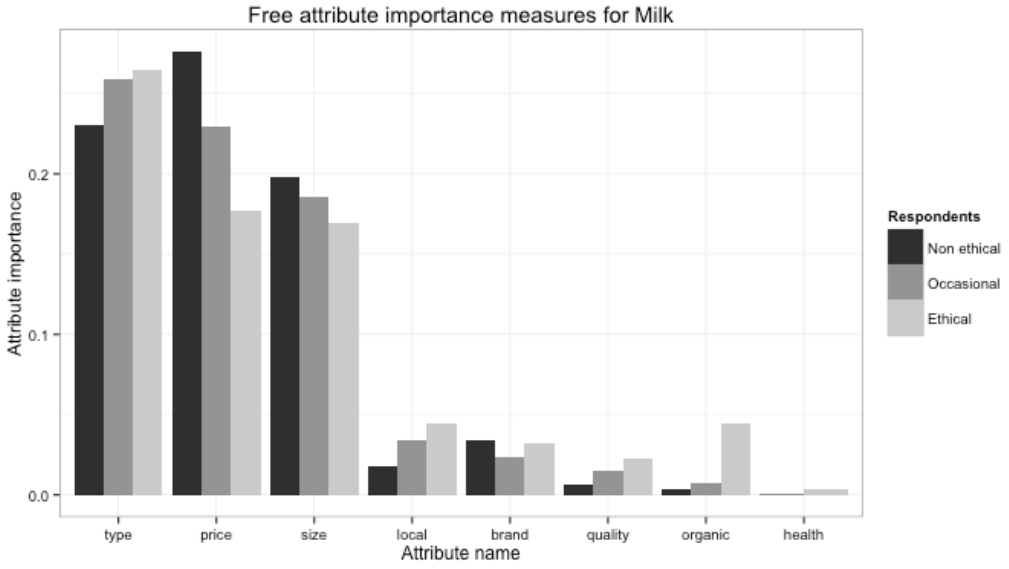


Figure 11: Attribute importance for milk elicited through free elicitation. The frequency with which an attribute was mentioned is taken as a measure of how important it is.

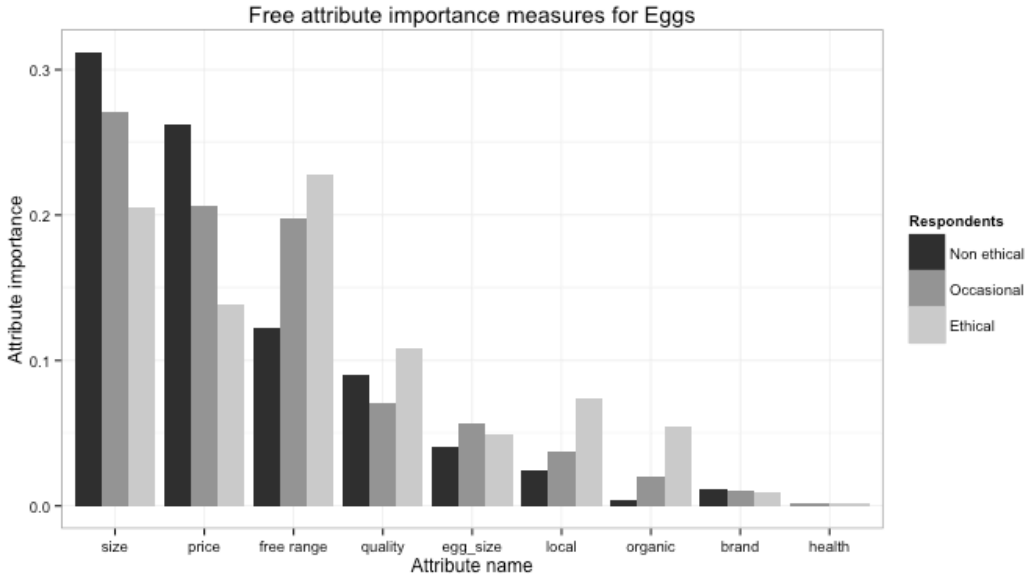


Figure 12: Attribute importance for eggs elicited through free elicitation. The frequency with which an attribute was mentioned is taken as a measure of how important it is.

For milk, the attribute ranking, per group, was similar to the prompted stated preferences ordering, tau = 0.714, 0.714, 0.809; $p < .05$ respectively for ethical, occasional and non-ethical consumers. This was not the case for eggs, where none of the rank correlations were significant. The attributes, and their importance, seems

therefore to be more accessible for milk, which is also rated as as most frequently purchased by consumers. As expected, ethical attributes such as organic, locally sourced and free-range were mentioned more frequently by ethical consumers who also mentioned price less often. Brand was again generally unimportant, especially for eggs, presumably because brands are not prominent in advertising and marketing for this product category. Interesting additional attributes, not present in the other direct elicitation task include 'freshness', which was ranked in as 3rd or 4th most important for milk, and 4th or 5th for eggs, depending on the consumer group.

Revealed preferences: Attribute importance

A multinomial logistic regression was fitted to the results of the DCE for the three respondent groups. Louviere and Islam's (2008) method was used to extract attribute importance by calculating each attribute's contribution to the model's log likelihood. The results are presented in Figure 13 and Figure 14. Again, the results are product dependent.

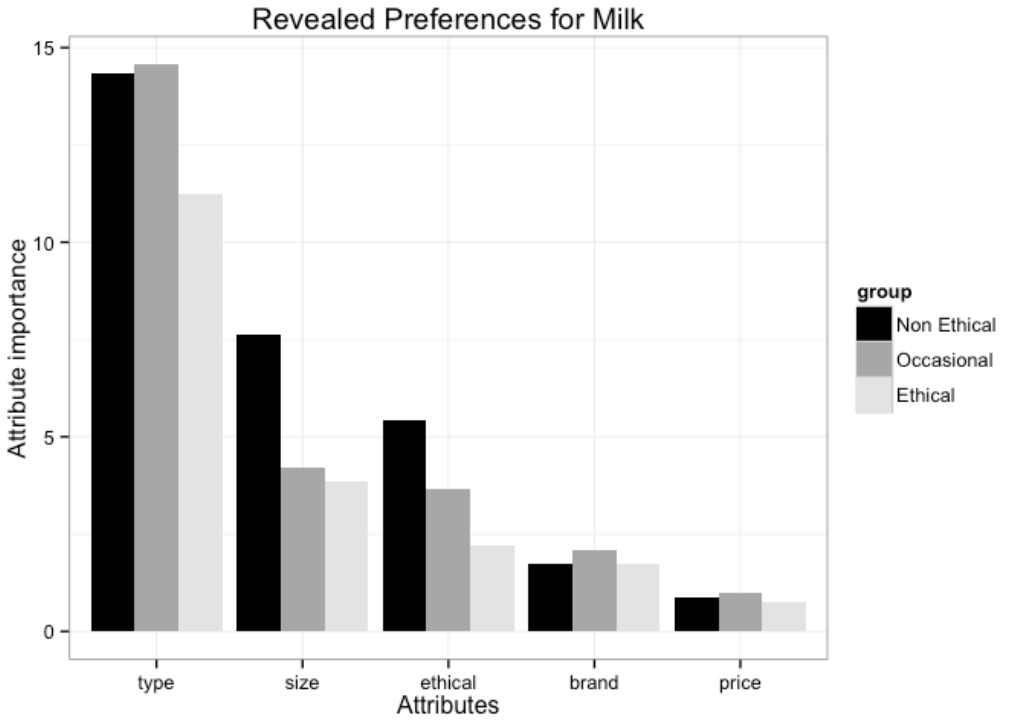


Figure 13: Revealed attribute importance for milk. There are less attributes present here since only attributes which are present in the DCE could be included.

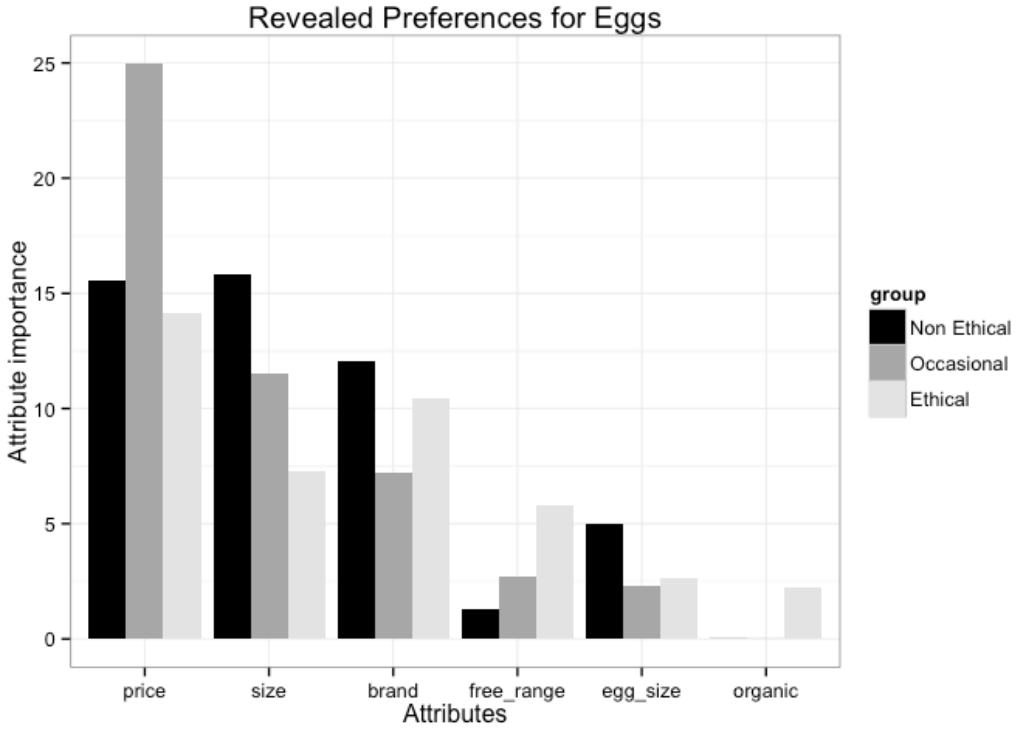


Figure 14: Revealed attribute importance for eggs. There are less attributes present here since only attributes which are present in the DCE could be included.

For milk, unlike what the respondents stated directly, price was the least important. Although this could be due to its relatively low variance (£0.22 for the un-normalised price, against £0.36 for eggs). Since the prices were all taken from real historical prices of milks the conclusion still holds: in the supermarket setting, within the range of prices displayed, consumers' behaviour is not highly affected by monetary concerns. All three groups showed the same behaviour with type and size as the most important attributes. As for the case of directly elicited preferences, brand did not have a strong effect on decisions. Although organic concerns seem to guide behaviour somewhat, as the sign of the coefficient within the regression model is negative, it is actually aversion to organic milks which drives behaviour. This is the case even for ethical consumers, as there is no discernable difference between the attribute importance ordering for the three groups. For eggs, price was the most important attribute, even for ethical consumers who did not report it so. Similarly, free range concerns were shown to be low for all three groups despite their high directly elicited preference. Brand is revealed as more important, especially for ethical consumers, which could point towards a halo effect of how ethical a brand is perceived to be.

A comparison between revealed and stated preferences, for all groups, confirmed the existence of the attitude behaviour gap. None of the correlations, for ethical, occasional, and non-ethical consumers were significant. Taking all consumers together and comparing between their directly and indirectly elicited preferences was also non-significant. Even for milk, for which directly elicited preferences were more stable across elicitation methods, there was no significant correlation between direct and indirect preferences.

	Ethical	Occasional	Non-ethical	All
Milk	$r_s = 0.4$, $p = .52$	$r_s = 0.3$, $p = .68$	$r_s = -0.1$, $p = .95$	$r_s = 0.3$, $p = .92$
Eggs	$r_s = 0.09$, $p = .91$	$r_s = 0.486$, $p = .36$	$r_s = 0.714$, $p = .14$	$r_s = 0.314$, $p = .56$

Table 9: Spearman correlation between stated and revealed preferences for all consumer groups and products. None of the correlations are significant, clearly indicating an attitude behaviour gap.

General Discussion

This study aimed to explore the consistency of directly elicited preferences for ethical, less ethical consumers. Additionally, it aimed to compare these directly elicited preference with indirectly elicited preferences to establish how far any of the groups had introspective access to their preferences. In doing so, it highlighted the attitude-behaviour gap for familiar, low involvement products such as milk and eggs. Segmenting the respondents by how often they report purchasing ethical products has yielded three groups: ethical consumers, occasionally ethical consumers and non-ethical consumers. To some extent, the direct preference analysis, and attitudinal questions, have shown the expected variations whereby non-ethical consumers rate price as the most expensive attribute guiding their decisions. Ethical consumers differ from others in that they rate health, organic and ethical concerns more highly.

The analysis of free responses revealed an association between direct and freely elicited preferences for milk, but not for eggs. Milk, which was more frequently brought, showed a higher stability of directly elicited preferences between methods, which hints at an increased introspective access. This was confirmed in a separate analysis of the same preference data using predictive models of choice (in Chapter 3). In this analysis, participants' hypothetical choices in the DCE as well as general consumers' choices from the supermarket scanner data were predicted using the directly elicited preferences (using both Anchored Scales and Constant Sum Scales). The predictive models based on these preferences for milk predicted actual

choices better than the ones for eggs. Still, further research is required before a strong link between frequency of purchase and greater introspective access is validated. The free elicitation results also echoed the relative unimportance of ethical attributes. Even ethical consumers mentioned attributes such as organic, ethical or healthy very infrequently, especially for milk.

Looking more closely at these directly elicited preferences within the groups, and quantifying each individual's consistency shows us that the self-reported ethical consumers are in general less consistent across the different direct elicitation methods. This is surprising since we would expect consumers who actively state their ethical intentions to have them more available and therefore to show more consistency between different direct preference elicitation methods. Another counter-intuitive and unexpected finding is that the inconsistency is more pronounced for milk, which on all other measures indicated increased introspective access. Further work is required in order to fully understand and explain this effect.

In addition to looking at individual consistency within the groups, it was also informative to observe more general preference homogeneity within them. This analysis showed that, as a group, ethical consumers had much lower preference homogeneity. In addition to being more homogeneous in terms of preferences, the two other consumer groups were more similar between themselves. To understand the low preference homogeneity for the ethical consumers, it is useful to go back to how the groups were formed. Consumers were classified into groups based on their self-reported propensity of buying "ethical products". The exact meaning of "ethical" was left purposefully open to interpretation. Some consumers therefore, might bear more importance on environmental issues, fair trade and social issues or animal cruelty. Given this plurality of ethical concerns, the heterogeneity of ethical consumers' preferences is less surprising. The practical implications of the lack of homogeneity is that marketers should be wary of appealing to ethical consumers as a single group, as their directly elicited preferences diverge significantly when measured for specific products. It is therefore preferable to categorise them in a more granular fashion.

All three groups exhibited an attitude behaviour gap, identified by the lack of correlation between their directly and indirectly elicited preferences. As this was

shown for all groups and for both product groups, it seems like this gap is not necessarily driven by ethical concerns, but is a more general phenomenon. This also eliminates social desirability as a possible source for this effect, since consumers who stated not being ethical should not feel need to express more sociably desirable ethical preferences. The revealed preference analysis also uncovered only very minimal difference between the attribute importance ordering of three consumer groups. This suggests, regardless of any difference in directly elicited preferences, a strong similarity in indirect, choice derived, preferences. Thus, consumers might self-report different attribute importance orderings, in line with their self-expressed ethical purchasing habits, but their behaviour reveals a very similar set of preferences.

Some of the differences between directly and indirectly elicited preferences can be explained by the variations present in the data. Indeed, price might strongly affect milk choice, but the variations in the available prices of milk is too small to have a large effect. Consumers might have been more affected by price differences if these had been more variable in their choice context. Since the prices both in the supermarket and the experiment have a relatively low variation, they might all fall within acceptable bounds and therefore not come out as strong drivers of behaviour in our analysis.

Conclusion and Future research

The current study showed that there is limited ethical concerns when buying milk or eggs, even for self-labelled ethical consumers. The variations of directly elicited preferences between the different groups of consumers does show that ethical consumers rate health, organic and ethical concerns more than the occasionally or non-ethical consumers. The indirectly elicited preferences of the three groups are quite different from their directly elicited ones, thus identifying an attitude behaviour gap. Interestingly, this gap is not specific to ethical dimensions, and is therefore less likely to be due to a social desirability bias.

Although there are only two product classes taken into consideration in the current study, it contributes to the understanding of ethical consumerism in two important ways. Firstly, the large scale comparison of direct and revealed elicitation methods, and the validation of the latter using real supermarket choice data, allowed

for a rigorous differentiation of the attitudes of different groups of consumers. This in turn led to the quantitative identification of the non-ethically lead attitude-behaviour gap. Additionally, the lack of strong differentiation between the indirectly elicited preferences of ethical and non-ethical consumers indicates similar choice behaviour for all groups when purchasing milk and eggs. Although they should be more deliberative and conscious in their decisions (Crane & Matten, 2005), ethical consumers did not show any increased introspective access to their preferences than the other two groups. Directly and indirectly elicited preferences were not aligned for any group, and the differences were not limited to ethical dimensions. This highlights the general limitation that people have in correctly expressing the drivers of their behaviour.

From a more applied angle, ethical consumers have been shown to be a group with very heterogeneous directly elicited. This suggests the need for a more fine-grained and personalised marketing strategy which should be tailored to individual consumers and their specific concerns such as health, fair trade, etc. Now that the external validity of the study has been established, further research could use different product categories such as vegetables which are more aligned with organic consumption, and are often bought outside of supermarkets. Building on the work presented in this chapter, these studies could also attempt to use a within-subjects design to compare directly and indirectly elicited preferences. This would enable an individual assessment of preference stability and identify the attitude behaviour gap at the individual level, leading to a greater understanding of ethical consumers.

Chapter 5. The malleability of direct and indirect preferences: the case of anti-American attitudes in urban Pakistan.

Abstract

This study compares the correspondence and malleability of directly and indirectly elicited attitudes, by taking American attitudes in Pakistan as an example. Attitudes are measured directly using a number of questions previously used by Gallup Pakistan and the Pew Global Attitudes Project, as well as indirectly using a discrete choice experiment. Respondents are first shown a pro-American advertisement with either high or low informational content, or a control advertisement for a washing powder. The pro-American advertisements have run in Pakistan between 2015 and 2016 by USAID and the American embassy in Karachi. The adverts did not have any effect on attitudes and general attitudes towards the USA were very negative. This is in line with historical results which put the USA's favourability in line with that of India's, regardless of the amount of advertisements, aid and development projects which it funded in Pakistan. Preferences are often different when measured directly and indirectly, yet in this case attitudes measured by both methodologies are well aligned. This indicates that these attitudes are introspectively accessible and validates the methodology used by the surveying projects such as the Pew Global Attitudes.

Introduction

Eliciting people's attitudes and preferences can be done using both direct or indirect methods. Oftentimes, preferences elicited in different ways diverge (Mueller et al., 2009; Van Ittersum, Pennings, Wansink, & van Trijp, 2007), resulting in the popularity of indirect methods for preference elicitation and choice prediction (Green & Srinivasan, 1990; Netzer et al., 2008). Such malleability has also been identified for attitudes which can be influenced by response mechanisms. In the following study, I propose to verify the agreement between directly and indirectly elicited political attitudes, thus uncovering if these are introspectively accessible. Additionally, I will explore the malleability of these directly and indirectly elicited attitudes to short term persuasion attempts, using anti-American attitudes in Pakistan as an example. In addition to contributing to the general discussion on the accessibility of attitudes, this work will also inform on the strength of anti-American attitudes in Pakistan and the effectiveness of current advertising campaigns on changing these attitudes. The rest of this paper is organized as follows: first I will give a little more background about preference elicitation as well as the mechanism that influence political attitudes. Following this, the survey-based experiment, run on a sample of participants in urban Pakistan will be presented. Finally, the results will be discussed along with conclusions and practical implications of attitude elicitation in developing countries.

Literature review

Preference and attitude elicitation methods are often categorized into direct and indirect methods. Direct elicitation, which accounts for much of the international attitudinal surveys (e.g. Pew Global Attitudes Project, 2013) assumes that respondents are able to correctly identify and report their attitudes. Direct elicitation questions are susceptible to a number of biases which may affect their validity. Response scale effects (Schwarz et al., 1985) and framing (Tversky & Kahneman, 1986), for example, can change the responses based on the context, wording or response mechanism of the questions. Often, respondents infer meaning from the context of questions (Schwarz & Oyserman, 2001), or even the response mechanism available which can change their responses and invalidate to some extent the elicited attitudes. In response to these biases, indirect elicitation methods attempt to deduce attitudes and preferences without requiring direct introspection. For example,

discrete choice experiments (DCE; McFadden, 1986) require respondents to repetitively state their preferences amongst a number of different alternatives. Preferences are then inferred indirectly for the different dimensions of the alternatives, by fitting a choice model such as a multinomial logistic regression. Other indirect elicitation methods attempt to access an attitude by using a different one as a proxy. For example, preference regarding the country of origin of products can be used to inform about attitude towards the country itself, as well as its reputation for quality and reliability (Abraham & Patro, 2014).

A distinction generally made with regard to both preferences and attitudes concerns whether they are implicit or explicit. Implicit attitudes, often measured through priming procedures and reaction time measures such as the implicit association test (Greenwald et al., 1998), are seen as being automatically activated and are inaccessible to introspection (Fazio & Olson, 2003). Explicit attitudes on the other hand are those which are uncovered though direct elicitation, in the absence of methodological biases such as the social desirability effect (DeMaio, 1984). Hence explicit attitudes are those which people can report (for example through questionnaires) while implicit ones are usually deduced from their behaviour. Both measures do not always correlate with each other (Nosek, 2007), and although both types of attitudes are generally believed to influence behaviour. Implicit attitudes are seen as better predictors of behaviour when cognitive load is high and actions are less deliberative (Friese et al., 2006). In the context of attitudes towards a country, implicit attitudes have been shown to be moderately correlated to explicit ones, showing that although both measures of attitudes are connected, they are distinct (Maier et al., 2015). In the absence of explicit attitudes, implicit ones have also been shown to be useful in predicting choices. For example, Arcuri et al. (2008) have shown that the choice of voters who stated being undecided, could be predicted by their implicit associations. Indirect preference elicitation methods, while accessing to some extent implicit associations do not necessarily directly measure these implicit attitudes. For instance, DCE which rely on deducing preferences from choices, may be done deliberatively with participants carefully weighting the different attributes of each option, or automatically by highlighting unconscious drivers behind choices (Mueller et al., 2009). Although other indirect questions, such as country of origin or country to which to emigrate to, rely to some extent on implicit associations these

are not directly measured by them. Hence, in our current study, there is no clear cut differentiation between both types of preferences: the direct elicitation methods uncover explicit attitudes, but the indirect measures uncover both explicit and implicit ones.

Attitudes and preferences, be they implicit or explicit, can be influenced and modified (Dasgupta & Greenwald, 2001; Petty & Wegener, 1998). There is evidence that information and news sources change political attitudes (DellaVigna & Kaplan, 2007), as can development programs (Beath, Christia, & Enikolopov, 2011). Political attitudes are also shaped by the repetition of information (DeMarzo, Vayanos, & Zwiebel, 2003) possibly via mechanisms such as the availability bias (Tversky & Kahneman, 1973). This is why, similarly to the consumer context, political attitudes can also be influenced by advertising. Political adverts have shown to have an effect on voter's knowledge levels the positive evaluation of candidates (Atkin & Heald, 1976) and voter intentions (Atkin, Bowen, Nayman, & Sheinkopf, 1973). General advertisements attempt to influence people's attitudes towards a product. This does not always involve the transmission of information. For example Resnik and Stern (1977) came to the conclusion that only around 50% of the TV advertisements had any informational content. Using the same methodology, Abernethy and Franke (1996) showed that 70% of TV advertisements have one or more informational cue, which drops to 33% for two or more cues. Yet advertisement information content is the single factor most correlated with overall value (Ducoffe, 1995), and the best predictor of brand attitudes (Aaker & Stayman, 1990). Reverting back to the context of general attitudes, it is interesting to see if these results hold: i.e. are informationally rich advertisements better at influencing attitudes than informationally poor ones?

Pakistan, and more specifically American sentiment in Pakistan, was chosen as the target this research for a number of reasons. First of all, as with other majoritarily Muslim countries, Pakistan has a generally unfavourable view of the USA (Pew Global Attitudes Project, 2013). As well as being positively correlated with incidences of terrorism (Krueger & Malecková, 2009), such high anti-American sentiment allows for the identification of a potentially strong effect in improving these attitudes. Additionally, Pakistan is considered a partner of the US in the war on terror (Delavande & Zafar, 2015) and is a receiver of significant amount of aid by

the US government (USAID, 2015). The identification and improvement of Pakistani attitudes towards the US is therefore important in advancing this political agenda and ensuring an impactful spending of foreign aid funds.

In the following study, I propose to verify the agreement between directly and indirectly elicited political attitudes, thus uncovering if these are introspectively accessible. Additionally, I will explore the malleability of directly and indirectly elicited attitudes to short term persuasion attempts, using anti-American attitudes in Pakistan as an example. In addition to contributing to the general discussion on the introspective accessibility of attitudes, this work will also inform on the strength of anti-American attitudes in Pakistan and the efficacy of informative and non-informative adverts in changing them. This deep exploration of the current attitudes towards the US will also be of interest to policy makers who are investing in improving the USA's image oversea.

Method

A sample of 1265 Pakistanis from urban Pakistan were surveyed as part of a weekly TV rating data collection. The survey was conducted face to face by Gallup professionals, and presented after the usual data collecting activities. The survey was translated from English into Urdu by professional translators, and back-translated by Gallup Pakistan. The questions and answers were read by the interviewer in Urdu and recorded on an android device using the ODK collect framework. Due to the limitations of ODK collect, the order of the answers was not completely random for each question. Instead four arbitrary orderings were created for each question, ensuring that USA was present at the beginning, end and in the middle.

The survey started by directly eliciting the participants' overall opinion of a number of countries (China, India, Pakistan, Saudi Arabia, Russia, USA, and United Kingdom) using a 0-10 scale. Participant then watched one of three short video advertisement[6]: an advertisement for a local detergent brand (control), an

[6] The three advertisements are available on YouTube on the following links (last accessed in November 2017):

- Control: (detergent): https://www.youtube.com/watch?v=_v8_QUkK6Ak
- Informational (USAID): https://www.youtube.com/watch?v=DKhsRIUJg2g
- Non-informational (embassy): https://www.youtube.com/watch?v=v7RuuyDHpbo

advertisement for USAID (informational advertisement condition) or a short video by the US embassy in Karachi showing their American staff discovering Pakistani mangoes (non-informational condition). Previous exposure as well as trust and likability of the messages were then measured. Subsequently, participants were introduced to a Discrete Choice Experiment about polio educational programs. Their attitude towards polio immunization programs was directly measured before they were faced with six random pairwise choices between two educational programs. Educational rather than direct immunization programs were chosen for the DCE in case there were religious or ideological opposition towards immunization in general. The variable attributes in the DCE were the country sponsoring the program, the price of the intervention (Rs 37, Rs 51, Rs 86 lakh which is equivalent to around 35000, 48000 and 82000 USD) and which month the intervention was to start (February or October, two months which do not have any Pakistani holidays associated with them). The months were used to create more variety and make it less obvious that the attribute with most importance for the research is the sponsoring country.

After the DCE, the overall opinion of the countries was once again directly measured and other direct and indirect measures of country preference were elicited. Finally, a short demographics section concluded the survey. For more details on the specific questions that were asked see Table 10, the full wordings of the survey is also available in Appendix 2. Although the preference for a number of countries was elicited, the main concern of this survey was specifically for the USA.

Dimension measured	Elicitation method	Wording excerpt	Response scale	Countries	Reference
Overall opinion	Direct	".. overall opinion of the following countries"	11 point scale (0-10)	China, India, Pakistan, Saudi Arabia, Russia, USA, and UK	(Delavande & Zafar, 2015; Pew Global Attitudes Project, 2013)
Indirect country preference	Indirect	Polio educational program DCE		China, India, Pakistan, Saudi Arabia, Russia, USA, and UK	
Indirect country of origin	Indirect	Choice of cars	chose country of origin of a gift car	Japan, USA, UK, Russia	(Gallup Pakistan, 2016)
Overall trust	Direct	"… overall trust in the following countries:"	11 point scale (0-10)	China, India, Pakistan, Saudi Arabia, Russia, USA, and UK	(Gallup Pakistan, 2016)
Interest of Pakistan	Direct	"…the following countries take into account interests of […] Pakistan?"	4 point scale from "Not at all" to "A great Deal"	Russia, UK, USA, China	(Pew Global Attitudes Project, 2013)
Partner of Pakistan	Direct	"…a partner […], an enemy of Pakistan, or neither?"	More of a partner, more of an enemy, Neither	Russia, UK, USA, China	(Pew Global Attitudes Project, 2013)
Influence in world	Direct	"…having a […] positive or […] negative influence in the world?"	5 point Likert scale from "Very Negative" to "Very Positive"	Russia, UK, USA, China	(Gallup Pakistan, 2016)
Indirect gratitude	Indirect	"…anonymously thank for their assistance to Pakistan."	check all that apply	China, UK, USA, Saudi Arabia, None	Adapted from Bursztyn et al. (2014)
Emigration	Indirect	"… would you like to move permanently to another country […]?"		Free response	(Gallup Pakistan, 2016)

Table 10: Summary of the elicitation methods used.

Results

Participants

The 1265 urban Pakistanis surveyed were part of Gallup Pakistan's television rating panel. Only 482 of these gave their consent to be part of the experiment. Due to logistical barriers, it is impossible to know de demographics of the rest of the sample. Respondents were randomly distributed amongst the three conditions with 158, 165 and 159 participants seeing the Control, Informational and Non-informational advertisement respectively. The responding sample was composed of 47.5% females. 52% of all respondents have some university education while a 35% had some secondary or tertiary education. The age of respondents ranged from 16 to 68 with a median age of 32 (and a mean of 35). Only 22% of the sample consumed news in English.

Advertisements

The three advertisements differed in their familiarity. Respondents were more familiar with the control condition (the washing machine advert) with 77% stating that they have seen it before, compared to 59% for the informational USAID advert, and 37% for the non-informational US embassy video (X-squared (2) = 50.8934 $p < .001$). Prior to conducting a MANOVA, the correlation between liking, understanding and trusting the message of the videos was calculated. The values ranged between 0.66 to 0.77, showing moderate to high correlation between the measures.

A multivariate analysis of variance was then conducted to check if the differences in opinions about the videos were statistically significant (Pillais' Trace = .053, $F(1, 480) = 8.967$, $p < .001$). The subsequent ANOVAs for each of the measures in turn confirmed that the liking measure [$F(1, 480) = 4.46$, $p= .035$] and the trust in the message [$F(1, 480) = 11.43$, $p < .0001$] were different amongst conditions, while the understanding of the video showed no significant differences [$F(1, 480) = 0.1123$, $p = .74$], despite the non-informational one being mostly in English. Post hoc comparisons using the Tukey HSD test indicated that the non-informational advert was more trusted than both the control ($M = 1.32$, $p = .0022$) and the informational advert ($M = 1.13$, $p = .01$).

Advert	Like	Understand	Trust the message
Control	6.21 (3.44)	6.6 (3.44)	5.15 (3.43)
Informational	6.63 (3.35)	6.96 (3.44)	5.35 (3.52)
Non-informational	7.02 (3.47)	6.47 (3.51)	6.48 (3.51)

Table 11: Ratings (and SD) of Liking, Understanding and Trusting the message of each advert. These are measured on a 10 point scale.

Direct elicitation methods

The individual measures of overall opinion of the USA before and after the viewing advertisement of the different countries did not differ significantly (see Table 12). Hence there is no identified overall change in attitude towards the US within participants. Similarly, the between-condition difference of post advertisement overall opinion of the USA was also not significant. All the other direct elicitation methods did also not show any significant difference in the attitudes towards the USA between conditions. These comparisons are summarized in Table 12.

		USA specific value					
Dimension		**Control**	**Informational**	**Non-info**	**Test**	**Statistic**	**P**
Overall opinion (post advert)		3.53[1]	3.19[1]	3.82[1]	ANOVA	F(1,480) = .1	.75
Overall trust		3.06[1]	3.32[1]	3.11[1]	ANOVA	F(1,480) = .024	.87
Interest of Pakistan		2.16[2]	2.18[2]	2.21[2]	Chi-square	$X^2(6)$ = 5.5916	.47
Partner of Pakistan[5]	Partner	21[3]	30[3]	27[3]	Chi-square	X^2 (4) = 2.9618	.56
	Enemy	97[3]	99[3]	89[3]			
	Neither	33[3]	30[3]	38[3]			
Positive Influence in the world[5]		2.33[4]	2.32[4]	2.49[4]	ANOVA	F(1,460) = 1.229	.27
Car choice		13[3]	14[3]	15[3]	Chi-square	$X^2(2)$ = 0.1429	.93
Indirect gratitude		24[3]	21[3]	25[3]	Chi-square	$X^2(2)$ = 0.3714	.83
Emigration		0[3]	2[3]	2[3]	none		
DCE		No difference, see Table 20 for details					

Table 12: Summary statistics for between-condition elicitation of American attitudes. None of the differences are significant, hence there is no evidence that advertisements modified any of these attitudes. Results from the DCE are reported seperately. 1. Values are based on a 10 point scale. 2. Values are based on a 4 point scale. 3. Values represent actual number of respondents. 4 Values are based on a 5 point scale. 5 Removed "Don't know" responses (less than 20 in total).

Updating of beliefs

Although the pre and post advertisement elicitation of overall opinion of the countries were separated by six questions and six binary choices in the DCE, there was still an anchoring effect of the initial overall opinion rating on the post-advertisement overall opinion rating. This could explain the lack of significant differences between the pre and post advertisement overall opinion measures. Indeed, in each condition, around 50% of respondents did not update their attitudes (see Table 13). Most of the respondents who did change their attitudes only did so by 1 point. A chi square text failed to identify any differences between conditions [X^2 (4) = 3.2169, p = .5222], confirming the lack of effect of the advertisement on respondents.

Condition	Unchanged	Positive Revision	Negative Revision
Control	54% (69%)	23% (18%)	22% (13%)
Informational	50% (70%)	27% (17%)	23% (13%)
Non-informational	53% (75%)	19% (10%)	28% (15%)
All conditions	52% (71%)	24% (15%)	23% (14%)

Table 13: Grouping of respondents per condition depending on how their attitudes were revised after viewing the advertisement. Values in brackets only take into account changes of more than one point.

Since there were no between-condition differences in attitudes, it was possible to compare perceptions about the USA in relation to the other countries, disregarding which advertisement was viewed. These overall measures are presented with the results of the questions which inspired them from the Pew Global Attitudes project as well as Gallup's past surveys.

For all direct elicitation methods, opinions of the USA were very low, only opinions of India were less favourable. Conversely, China, Saudi Arabia and Pakistan were viewed most positively. The measures of trust in each country were highly correlated with overall opinion (Table 14).

	India	USA	Russia	UK	Saudi Arabia	China	Pakistan
Overall Opinion	2.02 (2.68)	3.51 (3.33)	4.21 (2.9)	5.2 (2.84)	8.09 (2.68)	8.23 (2.73)	8.35 (2.46)
Trust	2.00 (2.68)	3.17 (3.2)	4.52 (3.02)	5.05 (2.68)	8.27 (2.4)	8.37 (2.62)	8.27 (2.59)
Correlation	0.71	0.59	0.61	0.47	0.55	0.63	0.59
Favourability (Delavande & Zafar, 2015)	2.02 (2.48)	2.58 (2.99)		4.54 (2.64)	7.88 (2.49)	6.91 (2.31)	3.23 (2.96)

Table 14: Overall opinion, measured before the advertisement, and trust, measured after the advertisement, for all countries The correlations between the two measures are reported, as well as the data from Delavande & Zafar (2015) to a similar question. All are measured on a 10-point scale, apart for the correlation coefficient. Mean is reported in each cell with standard deviation in brackets.

The USA was also seen as not taking the interest of Pakistan into account when making international policy decisions, as can be seen in Table 15. Again, China was seen to be well-aligned with Pakistan, followed somewhat distantly by the UK and Russia. Comparing the results with the data from the Pew Global Attitudes Project (2013) we can see that the latter is more skewed towards the lower end of the scale. A chi square test shows a significant difference in how the USA is rated ($X^2(3) = 61.4661$, $p < .001$), with 31% of respondents currently believing that the USA does not take into account Pakistani interests at all, compared with 54%.

In making international policy decisions, to what extent do you think the following countries take into account interests of countries like Pakistan?

	Great deal	Fair amount	Not too much	Not at all
UK	14%	47%	30%	9%
Russia	10%	43%	37%	10%
China	66%	27%	5%	1%
USA	10%	30%	29%	31%
USA (Pew Global Attitudes Project, 2013)	4%	18%	25%	54%

Table 15: Perception of the USA, UK, Russia and China's safeguarding of Pakistani interests.

In general, the USA was also viewed as more of an enemy of Pakistan, compared to the UK, Russia and China (see Table 16). This is consistent with the overall anti-american sentiment which emerged from the responses to the survey, and also quite similar to Pew's (2013) results, although much fewer respondents rated the USA as a partner in 2013.

Overall, do you think the following countries are more of a partner of Pakistan, more an enemy of Pakistan, or neither?

	Don't know	Neither	Enemy	Partner
UK	5%	39%	32%	25%
Russia	7%	49%	24%	20%
China	4%	15%	11%	71%
USA	4%	21%	59%	16%
USA (Pew Global Attitudes Project, 2013)	17%	13%	62%	8%

Table 16: Proportion of responses to the perception of countries as more of an Enemy or more of a Partner. Note that the USA is seen more as an Enemy, while Russia is predominantly perceived as neutral.

Finally, the USA's influence in the world was also seen as overall negative with 59% of respondents rating it either Very Negative or Mainly Negative. This is also in line with the results of the 2010 Gallup poll in which 50% of urban respondents rated America's influence as Negative (Table 17).

Please tell me if you think each of the following countries is having a mainly positive or mainly negative influence in the world?

	Don't Know	Very Negative	Mainly Negative	Neither	Mainly Positive	Very Positive	Average (1-5 scale)
UK	3%	6%	19%	35%	26%	11%	4.07[1]
China	2%	2%	3%	8%	25%	60%	5.32[1]
Russia	5%	9%	22%	37%	21%	6%	3.80[1]
USA	4%	31%	28%	15%	14%	8%	3.28[1]
USA 2010 (Gallup Pakistan, 2016)	9%	50%		27%	15%		

Table 17: Perceptions of the USA's, Russia's the UK's and China's influence in the world. Note that the Gallup 2010 results, presented in the 2016 report, were measured on a different 5 point scale: Don't know, Negative, Neither, Depends, Positive. We combined Neither and Depends to make it comparable. 1: P value for F test for the equality of means across countries is lower than 0.005 for all countries except USA-Russia.

Indirect elicitation methods

Overall, the indirect measurements also failed to show a significant difference between conditions. For the choice of cars task, the majority of respondents chose a Japanese car (more than 110 in each condition). American, British and Russian cars accounted for less than 12 % of choices in each condition (see Table 18). A chi-square test confirmed the lack of significant difference [x-square(2) = 0.14, p = .93] amongst the number of respondents who chose an American car, after seeing the different advertisements. Overall American cars were preferred to Russian ones, but only slightly. These values are quite different from the 2015 Gallup polls (Gallup Pakistan, 2016) in which Japanese cars had a much lower share of the choices (48%) while the USA cars were chosen by 20-25% of respondents, higher than both Russian and British cars. This difference could be due to sampling differences, as the current study used less than half of Gallup's sample.

Suppose you get a car as a reward, and four cars of the same size are placed in front of you. Which country's car would you like to have?

	Japan	Russia	UK	USA
Control	75%	9%	8%	9%
Informational	71%	9%	11 %	9%
Non-informational	74%	6%	10%	10%
All Conditions	73%	8%	10%	9%
Gallup May 2015	48%	7%	16%	25%

Table 18: Car choices for each of the groups, expressed as a market share percentage. As the question has been posed in past Gallup attitude surveys, the results are also included in the last row.

Only 4 respondents, across all the conditions, have mentioned that they are interested in emigrating to the USA while the vast majority (82%) stated that they would prefer to stay in Pakistan. More generally, there were no differences in the proportion of people wishing to emigrate outside of Pakistan regardless of if they were primed to think about their families first.

Similarly, the number of respondents who anonymously thanked the USA for sending aid was much lower than for the other countries, especially Saudi Arabia and

China who had gratitude expressed by between 36 and 42% of respondents (see Table 19). This pattern was similar to the one uncovered by the direct elicitation questions.

A number of countries have sent aid to Pakistan over the years. Please check which country you would be happy for us anonymously thank for their assistance to Pakistan.

	USA	UK	Saudi Arabia	China
Control	10%	15%	37%	37%
Informational	8%	11%	39%	42%
Non-informational	10%	12%	37%	42%
All Conditions	9%	13%	38%	40%

Table 19: Anonymous gratitude expressed for different aid providing countries.

Discrete Choice Experiment

The Discrete Choice Experiment was designed to give an average importance measure across participants in each group, as opposed to one measure per respondent. In order to ensure that participants' choices reflected their preferences, only choices of participants who did not say that they were against polio vaccination programs were taken into account. This resulted in 642, 673 and 703 choices for the three conditions respectively. In order to ensure that the difference in sample size does not affect the resulting models, 642 choices were repeatedly sampled from the second and third condition, and used in the construction of the models. The variable of interest was the relative importance of the country sponsoring the program. Three multinomial logistic regressions were fitted to the DCE using the mlogit package in the R programming language (Croissant, 2012).

The resulting models did not show much difference in the ordering of the countries, especially in the importance of the USA. The ordering, from most preferred to least preferred was: Saudi Arabia, China, Pakistan, Russia, UK, USA and India. The USA was generally indistinguishable from India as the least preferred country. Only in the informational condition was there a small significant ($p < .05$) effect, but this was not consistent when resampling the data and is therefore not considered a stable effect.

	Control	Informational	Non-informational
Intercept	-0.27** (0.09)	-0.34*** (0.09)	-0.28** (0.09)
Country Pakistan	1.55*** (0.26)	1.37*** (0.25)	1.04*** (0.23)
Country Russia	0.79*** (0.23)	0.37 (0.23)	0.12 (0.23)
Country Saudi Arabia	1.93*** (0.26)	1.53*** (0.24)	1.37*** (0.23)
Country China	1.66*** (0.25)	1.56*** (0.25)	1.33*** (0.24)
Country UK	0.76** (0.24)	0.80*** (0.22)	0.56* (0.23)
Country USA	0.18 (0.24)	0.42. (0.24)	0.03 (0.24)
Amount 51	0.15 (0.16)	0.11 (0.15)	-0.07 (0.15)
Amount 86	0.37* (0.16)	0.21 (0.15)	0.19 (0.15)
Month October	0.15 (0.12)	0.17 (0.12)	0.29* (0.12)

*Table 20: Three multinomial logistic regression models derived from the discrete choice experiment. One model was created per condition, taking only into account respondents who were not opposed to polio vaccination programs and keeping the number of choices used constant between all conditions. Betas for countries are calculated as compared to the base level of India. Standard deviations reported in brackets. Significance codes: . p < .1, * p < .05, ** p < .01, *** p < .001*

Discussion

In this study, I set out to compare the agreement between directly and indirectly elicited political attitudes, and uncover if such attitudes are introspectively accessible. I also explored the malleability of directly and indirectly elicited attitudes to short term persuasion attempts, using anti-American attitudes in Pakistan as an example. Attitudes in Pakistan are generally found to be negative, regardless of the high number of advertisement and development programs that are undertaken in the region by the USA. This was confirmed by the very low favourability scores on all measured dimensions and was in line with previous research (Delavande & Zafar, 2015; Gallup Pakistan, 2016; Pew Global Attitudes Project, 2013). The agreement between indirect and direct elicitation methods indicated that these attitudes are introspectively accessible. In addition to the lack of within participant differences (pre and post advertisement), there were no detected differences between conditions

on both the direct and indirect measures. Hence the informational and non-informational advertisements tested do not have an effect on anti-American sentiment, measured directly or indirectly, and do not affect short term attitude change. The USA has been sending aid to Pakistan for over 60 years (USAID, 2015) and despite this it is still perceived, implicitly and explicitly, as unfavourably as India, a country with which Pakistan has had a number of armed conflicts. This points towards the strength of the anti-American attitude present in the country, which does not seem to be strongly swayed by drawing attention to putative benefits of the USA-Pakistan relationship.

The two types of advertisement tested different in their content and source: one was produced by the USAID agency and the other by the American Consulate in Karachi. The latter was partly in English. The two advertisements did not differ in terms of their understanding, but respondents reported less trust for the message of the USAID video. The main difference between the content of adverts is that non-informational one did not contain any information about the USA's aid to Pakistan. This difference in information content and trust might partially explain why neither condition changed respondents' attitudes. The information provided by USAID could be discounted due to the lower levels of trust while the consulate's video did not provide any new information to counter the existing anti-American bias. Hence even if more trusted, there was no new information that participants needed to integrate.

Providing new information to respondents was expected to show some difference in attitudes, as was observed in Delavande & Zafar's (2015) study. Key differences between the two studies which might have limited the attitude revision include the presentation format of the information: in our study it was presented as an advertisement by USAID whereas Delavande & Zafar used small factual snippets of information such as amount of aid supplied by the USA. Additionally, the USAID advertisement did not have any specific amounts associated with how much aid is provided to Pakistan, instead it presented the diversity of projects which are funded (e.g. healthcare, energy...). Both the perception of the trustworthiness of the messenger (Brinol & Petty, 2009; Petty & Wegener, 1998), and the lack of precise information could have mitigated the persuasiveness of the advert. Finally, although the pre and post advertisement measure of overall opinion of the different countries

were separated by 12 questions, and the order of the countries was randomised in each case, there was still evidence of a strong anchoring effect. Indeed, more than 50% of respondents in each condition did not update their attitudes, and the majority of those who did update them did so by only 1 point on a 10-point scale.

Implicit and explicit attitudes do not always align but are nonetheless often positively correlated (Nosek, 2007). In our study, there was little difference between the ordering of the directly elicited overall favourability and the indirectly elicited country preferences. In both cases, the USA was among the two least preferred countries, along with India. While Saudi Arabia, China and Pakistan were the most favourably viewed. Indirect and directly elicited attitudes usually differ either when there is a conscious limitation in reliably expressing an attitude (e.g. DeMaio, 1984) or when the attitude is not accessible to introspection. As attitudes in Pakistan are generally overtly anti-American, and the interviewers were local it was less likely that respondents had to self-edit their attitudes towards the USA. Putting this together with the general lack of variation in self-reports pre and post advertisement, it appears that people have access to their attitudes towards the USA which is stable. Indirectly elicited attitudes are often more predictive of behaviour in domains such as interracial or intergroup behaviour (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). On the other hand, self-reports better predict consumer and political preferences. Our study validated the fact that American attitudes in Pakistan are closer to this latter domain, and can therefore be researched using straightforward self-reports as is currently done by surveying projects such as the Pew Global Attitudes Project or Gallup Polls.

External validity of the direct elicitation measures was given by the comparison of the results with previous survey questions used by the Pew Global Attitudes project (2013), Gallup Pakistan (2016) and Delavande & Zafar's (2015) study. Although respondents were only sampled from Gallup Pakistan's urban sample, and were therefore different from these previous studies, the results were generally very similar. For example, Delavande & Zafar's (2015) measure of overall favourability for the USA using a similar 10 points scale differs by only 1.07 points, and the ordering of the countries is similar. One significant difference between the studies is with regards to Pakistan, which has a much higher favourability score in our work. This can be explained by a small difference in the wording of the question

in our survey which was about Pakistan as a country as opposed to the Pakistani Government. Additionally comparison with the similar direct elicitation questions by the Pew Global Attitudes Project (2013) and Gallup Pakistan (2016) show very similar patterns to those uncovered in this study (see Table 15, Table 16 and Table 17). The slight differences observed between surveys could be due to a number of factors such as a change of attitudes between 2010 and 2016, or the nature of the samples used. Yet, although the Pew Global Attitudes Project (2013) used a mostly urban sample of 1201 adults across Pakistan, the Gallup Pakistan (Gallup Pakistan, 2016) sample was similar to the one used in this study. Hence it is more likely that these differences are indeed comparable and that the trends identified reflect an actual attitude change.

External validity for the comparable indirect methods is slightly more complex. Although the DCE results mirror the direct elicitation responses, which suggests that they are indeed measuring an accessible attitude, both the emigration and car choices are more difficult to interpret. For the former, not many respondents considered emigrating, and even less considered the USA as a possible destination. This is a risk with many free elicitation questions. The question regarding choice of cars was very skewed towards Japanese cars, even more than when the question was used by Gallup Pakistan (Gallup Pakistan, 2016) in 2015. It seems that the attitude towards a country is not well correlated with country of origin, and in this case reputation for quality trumps the effect of this general attitude. On the other hand, the anonymous expression of gratitude did mirror quite closely the direct elicitation responses. This is even more remarkable since there was no incentive to do so, unlike Bursztyn et al.'s study (2014) where a payment of around a fifth of a day's wages was conditional on the expression of anonymous gratitude to the USA. In that study, there were still around a quarter of participants who did not wish to express their thanks. The current work showed that even without incentives, and by allowing multiple countries to be anonymously thanked, the methodology can be used to elicit comparative attitudes towards a number of countries, consistent with both direct and other indirect methods.

The comparison with past questions used by other studies can also situate the results historically. Indeed, although in general attitudes towards the USA are still low, there is some evidence of increased positive attitudes towards the USA between

2010 and 2016. The results from Delavande & Zafar's (2015) study in show that American attitudes were lower on a 10 point scale when the data was collected in 2012, as were the perceptions measured by the Pew Global Attitudes Project (2013) in 2013 (see Table 15 and Table 16). Some of these differences can be due to the level of media access that our sample has, compared to the more general population, but the results suggest a weak upwards trend in the positive perceptions of the USA.

Another practical outcome from this research is the applicability of using indirect preference elicitation methods for eliciting attitudes in developing countries. Surveying companies rarely engage in large scale DCE in developing countries, whereas they do have the infrastructure to run large scale self-reports. This study has shown the feasibility of such studies using ODK (Brunette et al., 2013), an offline mobile data collection toolkit which is often used in developing countries. External validity of the method was given by the correspondence of the DCE derived country preferences to the ones elicited by the direct methods. Generally, large surveys in developing countries are conducted by a number of interviewers going door-to-door. Often these interviewers do not have access to an internet connection, and the technical solutions that they can use to aid data collection need to be offline. The hardware that they have at their disposition is also often limiting. Simple, technologically limited, solutions such as ODK are therefore well-adapted to this domain. Other indirect methods, such as the IAT, which rely on precise reaction time measurements are more difficult to apply in this setting. Thus, using DCEs or conjoint analysis techniques, such as the ones demonstrated here, can be a good compromise.

Conclusion and further work

In short, our conclusions indicate that directly and indirectly measured attitudes towards the USA in Pakistan are not malleable and are accessible to introspection. Both direct and indirect elicitation methods revealed similar anti-American attitudes. Additionally, the study detected no effect of short advertisements, regardless of their informational content, in changing attitudes. This could be due to the composition of the sample, the lack of trust of the information provided, or even the strength of the attitude itself. As there is evidence that new information can change attitudes (Delavande & Zafar, 2015; DellaVigna & Gentzkow, 2009), the main recommendation that can be taken from this study is to

add more specific information (such as amount of aid spent) into these advertisements in order to make them more persuasive. The general low opinion of the USA identified in this study, especially amongst the sector of the population most prone to be consuming international news on television, suggests that better tailored communication is required in order to improve the perception of Pakistanis towards the USA. The correspondence of indirect and directly elicited attitudes, which implies introspective access to these attitudes, validates the methodology used by attitudinal surveys such as the Pew Global Attitudes Project and Gallup Pakistan. Further work can also use implicit measures (see Maier et al., 2015 for example) to investigate to what extent these match with directly and indirectly elicited attitudes, and if adverts have an effect on these implicit associations.

Chapter 6: Discussion and Conclusion

The main aim of this thesis was to explore the limits of introspection, concentrating on preference elicitation methodology: namely the correspondence and divergence of direct and indirect elicitation methods. To achieve this aim, I carried out three studies, presented in Chapters 3, 4, and 5, the results of which will be summarised and discussed in light of previous research. The strength and limitations of these studies will also be presented, along with ideas for future studies. Finally, I will conclude with a short summary of the contributions of this research both to theory and practice.

Introspective access to habitual purchases

The first study focused on the context of habitual consumer purchases, and people's introspective access to their preferences for such products. First, I examined the consistency of the attribute importance ranking collected by three direct elicitation methods. Although there were some product-specific variations, with milk preferences being more stable than the preferences for the other two product groups, the direct elicitation methods did not result in a stable, differentiable, attribute importance ordering. The orderings were nevertheless strongly correlated, in line with previous research (Chrzan & Golovashkina, 2006; Louviere & Islam, 2008). For example, when testing six direct methods of preference elicitation for restaurant preferences, Chrzan and Golovashkina (2006) found a very high correlation amongst the resulting preferences. Louviere and Islam (2008) mirrored these results as applied to the choice of pizzas and packaged juices amongst students. The first study of this thesis replicated these findings for even more familiar and habitual decisions, using a large non-student sample. Unfortunately, although strongly correlated, the attribute orderings were not stable: even amongst the most important attributes there was no statistically significant order for all elicitation methods. In addition to comparing the correspondence of direct methods amongst themselves, Louviere and Islam (2008) also compared the correspondence between direct and indirect elicitation methods, showing that they were significantly less correlated. The results of my first study mirrored this for all product categories. This strengthens Louviere

and Islam's (2008) results by extending the number of product categories under consideration, increasing their familiarity and using a more general sample.

The second aim of this study concerned the comparison of the different elicitation methods in terms of how well they inform both experimental and non-experimental choices. In this way, direct and indirect elicitation methods were each used to parameterise two different types of models: a weighted additive model and a lexicographic one. Previous studies comparing predictive validity between direct and indirect elicitation methods often only use a weighted additive framework. Thus, Srinivasan (1988) elicited preferences from MBA students about job offers and used weighted additive models to predict their final acceptance of jobs at the end of the year. Although the results indicated that directly elicited preferences were slightly more predictive, the improvement was not statistically significant. Akaah and Korgaonkar (1983), also compared a number of direct and indirect elicitation methods on preferences for health maintenance organization plans (a the time, a new type of health insurance plans), using weighted additive models. Their results showed that indirect methods were more predictive than direct ones at predicting hypothetical choices. The first study presented in this thesis strongly supports these results, both with a much larger non-student sample and using less complex and more familiar products. In addition to validating these results on hypothetical choices, the study also extends it to actual consumer behaviour. In contrast, when moving to a lexicographic framework, the indirect elicitation method used was significantly worse than direct elicitation methods.

Generally, lexicographic models are parameterised through direct elicitation techniques. For example, Scheibehenne, Miesler and Todd (2007) compared weighted additive and lexicographic models of food choice, based on direct preferences. Their results show a slightly better performance of the more complex weighted additive model. Their study does provide evidence for lexicographic processing of food choices, as these models, which relied on less than two attributes on average, predicted food choices nearly as well as the weighted additive models which integrate the preferences of all the attributes. The study presented in Chapter 3 finds a similar pattern of results, with lexicographic models, based on direct elicitation, predicting nearly as well as the full weighted additive ones, based on

indirect elicitation. As stated previously, the performance of the directly elicited weighted additive models was much lower than the one parameterised by the indirect method. One way of adapting indirect elicitation methods to parameterise lexicographical models was investigated by Dieckmann, Dippold and Dietrich (2009). They compared two models derived from indirectly elicited preferences: a lexicographic model (Kohli & Jedidi, 2007; Yee et al., 2007) and a weighted additive one, and similarly concluded that the weighted additive model outperformed the lexicographic one. Although using ski jackets as the product category, this is in line with the results of my first study on frequently purchased products. Hence, although there is some evidence for lexicographic decisions, weighted additive models based on indirect elicitation methods such as discrete choice experiments have greater predictive validity.

This last point, especially as applied to familiar decisions, implies that people do not have much introspective access to their decisions. Indeed, they are unable to reliably report the drivers of their behaviour, even in a lexicographic framework which should be less cognitive taxing. As these models do not require absolute values, they can be more easily parameterised from directly elicited preferences. This is especially true if these preferences are not actually stored as absolute values (Vlaev et al., 2011). The unreliability of direct elicitation methods and the superiority of the discrete choice experiment indicate that deducing people's preferences from their behaviour is more reliable than asking them to report it. This confirms results outside of the context of preference elicitation (Kolar et al., 1996). For instance, Vazire (2010) arrived at similar conclusions when comparing people's self-judgements of a number of their personality traits, with others' judgements of these traits. Traits with high evaluativeness, were more accurately judged by others, much like preferences which are better identified through observing choices.

Introspective access and ethical purchasing

Building on the results of the first study, i.e. identifying the unreliability of directly elicited preferences, the first aim of the second study (presented in Chapter 4), was to explore the consistency of direct preferences for ethical, occasionally ethical and non-ethical consumers. The respondents were divided into three groups

depending on their self-reported frequency of buying ethical products. The demographics, and measured general attitudes were consistent with the literature on ethical consumers (e.g. Sachdeva et al., 2015; D. Shaw et al., 2005) which increased confidence in the correct classification of consumers. The stated preferences, much like the general attitudes, were different amongst the three groups. These preferences also differed in how consistent they were within the groups. Non-ethical and occasionally ethical consumers were the most consistent: i.e. their stated preferences were more homogeneous. Ethical consumers, on the other hand, showed much more variability in their directly elicited preferences. This showed both the plurality of ethical concerns and the limited effect of general ethical concerns on specific product categories. Health and environmental concerns, which are generally found to be important attributes for ethical consumers (Essoussi & Zahaf, 2009; Joshi & Rahman, 2015) were as relatively unimportant for milk and eggs.

The second aim of this paper was to explore the degree to which direct and indirect preferences converge for the three groups of consumers identified. It was hypothesised that consumers with strong ethical concerns would be more likely to be consistent in their directly and indirectly elicited preferences, as they are considered more deliberative shoppers (Crane & Matten, 2005). I therefore compared indirectly elicited and directly elicited preferences for each group. For milk, indirect preferences were similar across groups, indicating that although ethical consumers' directly elicited preferences were different, their actual behaviour was similar to occasional and non-ethical consumers. For all three groups, there was a difference between directly and indirectly elicited preferences, indicating the existence of an attitude behaviour gap. This is consistent with previous research (Boulstridge & Carrigan, 2000; Sheeran, 2002). Auger and Devinney (2007) used a similar design as this study to show that direct survey questions overstate ethical concerns, when compared to preferences elicited through a discrete choice experiment. This was replicated in the present study, but instead of using students and supporters of Amnesty International, I used a general sample of UK consumers. Although consumers generally report that for milk products price is very important, their choices reveal that it is the least important attribute. For eggs, on the other hand, price is the most important indirectly elicited attribute for all groups. This result partially supports the general notion that price is one of the main barriers for ethical

consumption (Carrigan & Attalla, 2001), but this seems to be dependent on the product category under investigation. In short, the results of the second study provided evidence that ethical consumers have the least homogeneous preferences, and like all other consumers, display an attitude behaviour gap. Although ethical consumers have different directly elicited preferences, their indirectly elicited ones are much more similar to occasionally ethical and non-ethical consumers, highlighting an attitude behaviour gap which exists for all consumers. As this gap is not necessarily driven by ethical concerns, it could be the manifestation of a much broader phenomenon, namely consumers' limited introspective access to their preferences.

Introspective access to political attitudes

Having identified people's limited introspective access to familiar purchases (in the first study) and the fact that identification as ethical consumers does not increase this access (in the second study), the aim of the final study was to explore these findings outside of the context of purchases. Hence, the first aim of this study was to verify the agreement between directly and indirectly elicited political attitudes in Pakistan. The results of the study showed a high level of agreement between directly and indirectly elicited preferences towards the USA. Some of the specific measures of attitudes have been adapted from large scale projects such as the Pew Global Attitudes, or Gallup Pakistan's surveys. This allowed checking the study for external validity. The strong correspondence between both types of elicited preferences confirmed the use of direct preferences for these large projects. Indeed, more complex indirect methods, such as the discrete choice experiment, did not add much information as the simpler direct elicitation methods already captured all the information. Thus, unlike product preferences, political attitudes, such as anti-american sentiment, are more accessible to introspection. The very high anti-american attitudes, which previous research has identified in Pakistan using both direct (Delavande & Zafar, 2015) and indirect (Bursztyn et al., 2014) elicitation methods, was reflected in the current study. Additionally, the indirect elicitation technique presented by Bursztyn et al. (2014) which involved asking participants to anonymously express their gratitude towards the USA in return for a monetary incentive (around a fifth of a day's wage) was replicated in a non-incentive aligned

way. The alignment between direct and indirect preferences which has been uncovered by this study can also be explained in light of the influence that indirectly elicited preferences have on undecided people's choices (Arcuri et al., 2008; Galdi et al., 2008). Arcuri et al. (2008) found that indirect preferences, measured through the implicit association test (Greenwald et al., 1998) can predict the voting behaviour of undecided voters. In the case of our study, people were not undecided about their sentiments towards the USA, hence their indirect as well as direct preferences were aligned, and accessible.

The second aim of this study was to measure the change in directly and indirectly elicited attitudes after exposure to positive advertisements. Two different types of advertisements were tested: one advert with high information content, and one with a low information content. Both types of measured preferences did not show any short term change after exposure to any of the advertisements. This result goes against Delavande and Zafar's (2015) findings that anti-American attitudes in Pakistan can be reduced when people are exposed to (positive) information. The lack of such an effect in my third study could be because of the difference in the nature of information provided to the participants. Indeed, while Delavande and Zafar's (2015) information was presented as precise numerical information regarding aid, the information presented in the informational advertisement that was used only gave general information regarding the scope of USAID projects in Pakistan. The trustworthiness of the source of the message (Brinol & Petty, 2009) could also mediate the effect of the advertisement, as the advertisement was rated as only moderately trusted.

Strength and limitations

The main strength of the research reported in this thesis is its methodological consistency in exploring the limits of introspection in various domains. Different methods for directly eliciting preferences are compared in all three studies with indirect elicitation methods, including a discrete choice experiment. The first study also ensured the validity of the non-incentive aligned discrete choice experiment, which was used throughout. All studies used reasonably large, non-student participants drawn from the general population, in the UK (for studies 1 and 2) and

in Pakistan (for study 3). The comparison with actual consumer supermarket choices, in the first two studies ensured that they were externally valid, and that the results were not just artefacts of the experimental method. This was achieved in the third study by ensuring an overlap in the direct and indirect methods used with previous commercial and academic work in Pakistan (Bursztyn et al., 2014; Delavande & Zafar, 2015; Pew Global Attitudes Project, 2013). The availability of the large supermarket scanner dataset also helped to ensure the familiarity of both the products themselves and the choice context. Similarly, the third study used real advertisements which have been showing in Pakistan to measure their effectiveness: i.e. advertisements which have been created with the aim of improving attitudes towards the USA. Finally, given this thesis' dual focus on theoretical advances as well as practical uses, the research is both relevant academically and can be applied directly by marketers or surveying companies.

Nevertheless, this study does have a number of limitations which should be acknowledged. Firstly, the complexity of the decision context and the need to match the context to the data provided by the supermarket scanner dataset, to ensure external validity, meant that the discrete choice experiment was not designed in a way to maximise its efficiency. As, even when unoptimised, the choice experiment outperformed the direct elicitation methods, this inefficiency does not take away from the validity of the study's results. Concerns about the unincentivised nature of the discrete choice experiment were also alleviated in a similar fashion: by comparing hypothetical choices to ones present in the supermarket dataset. Another important limitation, which also stems from the use of the scanner dataset for measuring external validity, is that the design of the choice experiment did not allow for the identification of individual preferences. Instead, indirect preferences were elicited for the whole population, or group for the second study. Further studies could validate these findings at the individual level, and extend the results to other product categories. In the case of the last study, extensions could investigate the effect of more specific information on the elicited attitudes, or attempt to replicate the findings in different countries. This will verify if directly and indirectly elicited political attitudes are still aligned and therefore introspectively accessible in countries with less anti-American sentiment.

Conclusion

Overall, this thesis presents evidence for the following points, regarding the limits of introspection and the methodology of preference elicitation:

- Directly and indirectly elicited preferences are not consistent for familiar often purchased products. This means that these preferences are not introspectively accessible, since there is a difference between what people say and what can be deduced from their choices.
- Ethical consumers do not have more access to their preferences than non-ethical or occasionally ethical ones. Although they state that ethical attributes are more important to them, and they are considered more deliberative shoppers, they do not exhibit more consistency between their directly and indirectly elicited preferences.
- Political attitudes, as measured by anti-American sentiment in Pakistan, exhibits strong coherence between direct and indirect preferences. This is indicative of the availability of these attitudes to introspection, and confirms the reliance of large country-wide surveys on direct elicitation for measuring political attitudes.
- Methodologically, weighted additive models, parameterised from indirect preference elicitation methods, are more predictive of people's choices than models parameterised from direct preference elicitation. Although the latter can perform adequately in some decision contexts, when using a lexicographic framework to predict habitual decisions for example, the unreliability of direct preference elicitation limits the usability of such models.

Although theoretically interesting, these conclusions are also directly relevant to researchers and marketers working in practice. Firstly, as stated previously, it justifies the reliance of researchers studying political attitudes such as pro- or anti-American sentiment on stated preference methodologies. In the same vein, it also promotes the continued use of indirect preference elicitation methods to derive preferences for products, even if these are familiar and habitually purchased. These can be used to understand preferences, but also to predict products' popularity and

market shares. In the context of habitual decisions, or other environments which promote non-weighted additive models, lexicographic models have been shown to give surprisingly good results, compared with their simplicity. The applicability of these models is limited by the unreliability of directly elicited preferences, which could be improved with the advent of novel direct elicitation techniques.

In light of this research, future work can extend the results of this thesis in a number of directions. Initially, the inclusion of different types of habitually purchased products could reveal if these results depend on how frequently such products are purchased, and if more deliberative purchases allow for introspective access. Using individual indirect preferences in a similar context would also be an interesting extension, as it would be able to quantify more precisely the difference between directly and indirectly elicited preferences at the individual level. This could also be used to test the extent to which individual consumers engage in lexicographic choice strategies, and if these are aligned with their stated preferences. Finally, the work on the introspective access to political preferences could also be diversified to countries with weaker anti-American sentiment. This could help establish how far political attitudes are introspectively accessible, when they these attitudes are less polarised.

Appendix 1. Additional materials and analysis for Chapter 3.

A1.1: Features and levels for all product groups

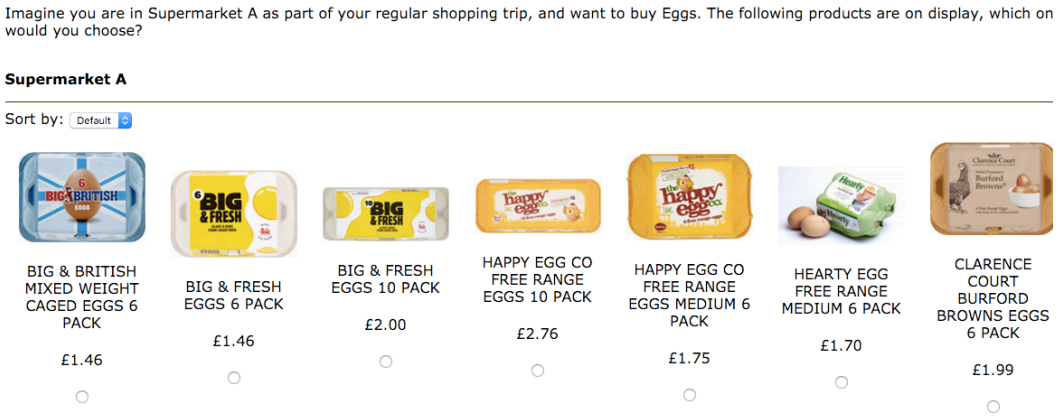
Milk:	Type	Price	Size		Brand	Organic
	Semi-skimmed Skimmed Whole 1% Fat Jersey		4 pints 2 litres 2 pints 1 litre 1 pint 6 pints 0.4 pints 225 ml		Supermarket-brand Cravendale Meadow Park Yeo Valley Flora Marybelle Calon Wen Acorn A2 Fresh n Low Lactofree	Y N
Egg:	Free-range	Price	Size	Egg-size	Brand	Organic
	Caged Free Range		6 eggs 9 eggs 10 eggs 12 eggs 15 eggs 18 eggs 30 eggs	Mixed Med Large Very Large	Supermarket-brand Big n British Big n Fresh Chippindale Clarence Court Eggs For Soldiers Happy Egg Hearty Natures Nest The Kent Egg	Y N
Tuna:	Type	Price	Quantity		Brand	
	In Water In Brine In Oil		1 large tin (200g) 1 medium tin (160g-185g) 1 small tin (120g) 1 very large tin (400g)		Supermarket brand John West Princes Royal Line The Reel Fish Co	

			2 medium tins (2x160g) 3 large tins (3x198g) 3 small tins (3x120g) 3 very small tins (3x60g-3x80g) 4 medium tins (4x160g-4x185g)			

Table 21: Breakdown of all the features and levels for each product group. Only the features in this table are used in the predictive models.

A1.2: Example wording of the different elicitation methods.

The discrete choice experiment task resembled an online grocery shop. Products are displayed on a grid sortable by price, size and type, and a button was presented at the bottom of the grid to submit the selection.



The wording of the Constant Sum Scales (CSS) preference elicitation method was as follows:

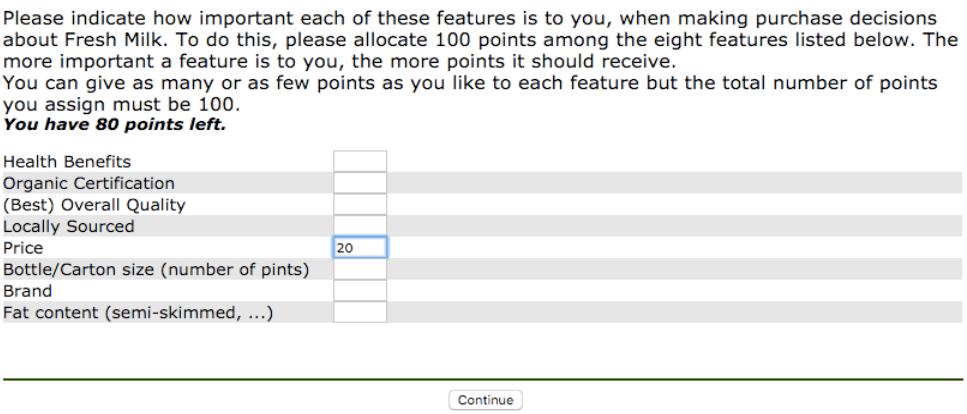


Figure 15: Example CSS question for attribute importance

The Anchored Scales question had two parts. Initially, respondents had to rank all of the attributes, or in the case of attribute levels, just specify their most and least important. Then they would rate all the other attributes or levels on that scale.

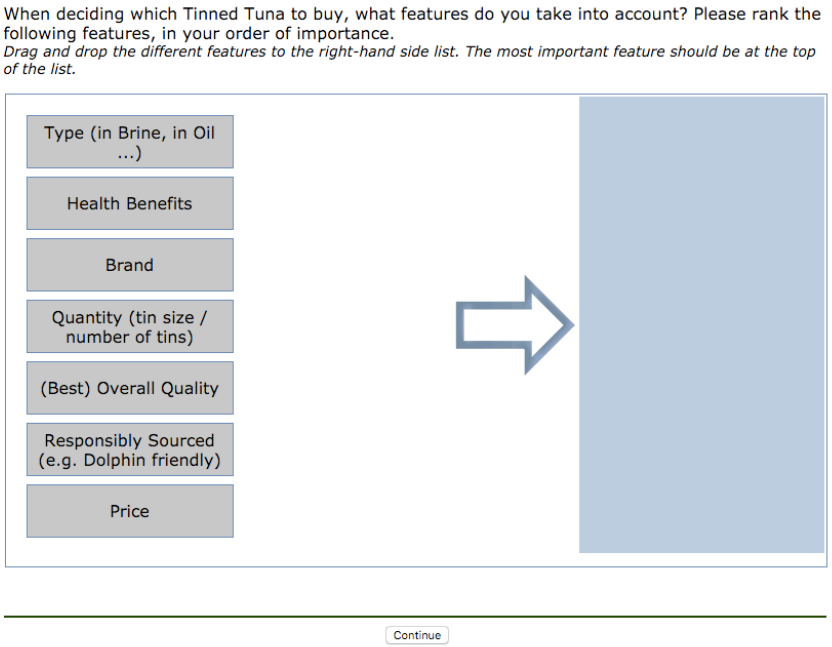


Figure 16: Example of the first step in the Anchored Scales methodology, where respondents are asked to rank the attributes.

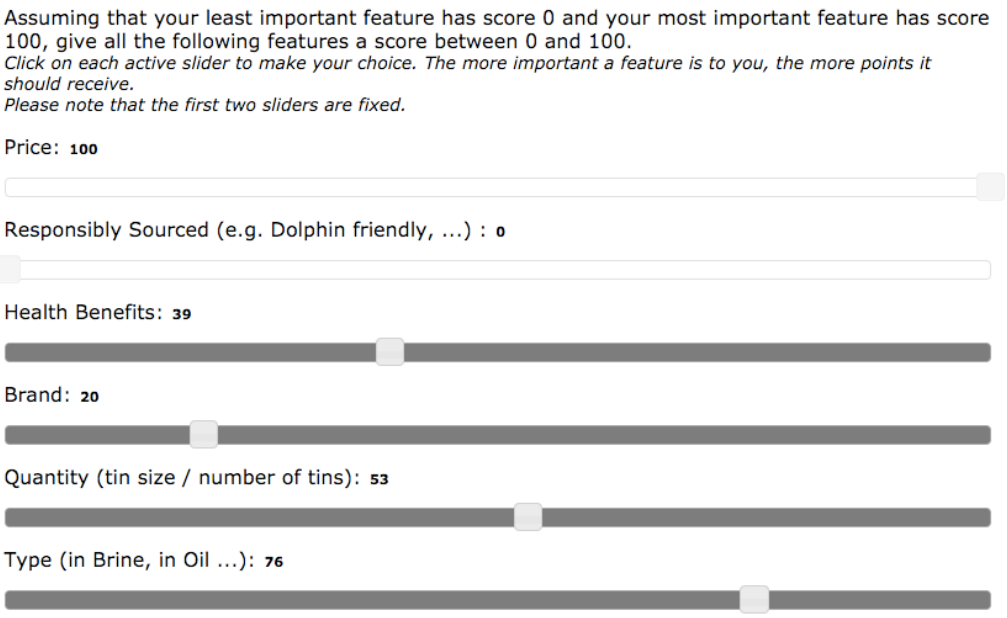


Figure 17: The rating step of the Anchored Scales methodology in which respondents rate all the attributes

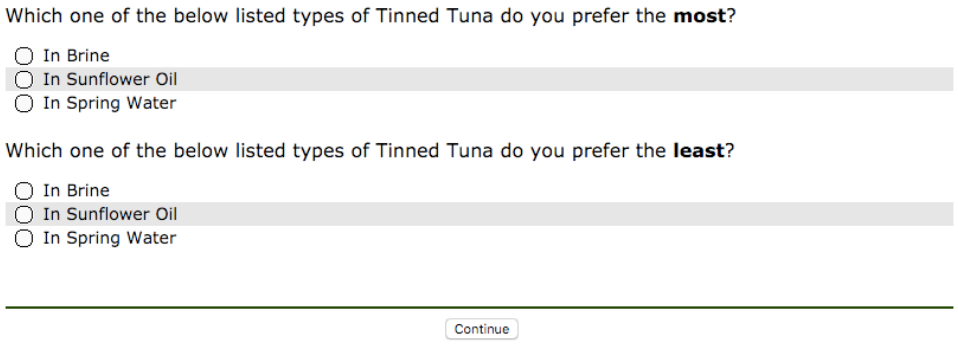


Figure 18: When moving to attribute levels, respondents in the AS methodology only have to specify their most and least preferred level.

A1.3 Details of additional analysis.

The results of running the hit rate comparison 30 times with models parameterized on only 50% of the respondents. As predicted the error bars are larger which translates to less robust models and hints at the instability of the respondent expressed preferences. The results for the milk products are an exception which signals either a high level of introspective access or preference homogeneity in the

population. On the other hand, the sharp decrease hit-rate for the CSS LEX eggs model points towards the lack of robustness of the directly elicited preferences.

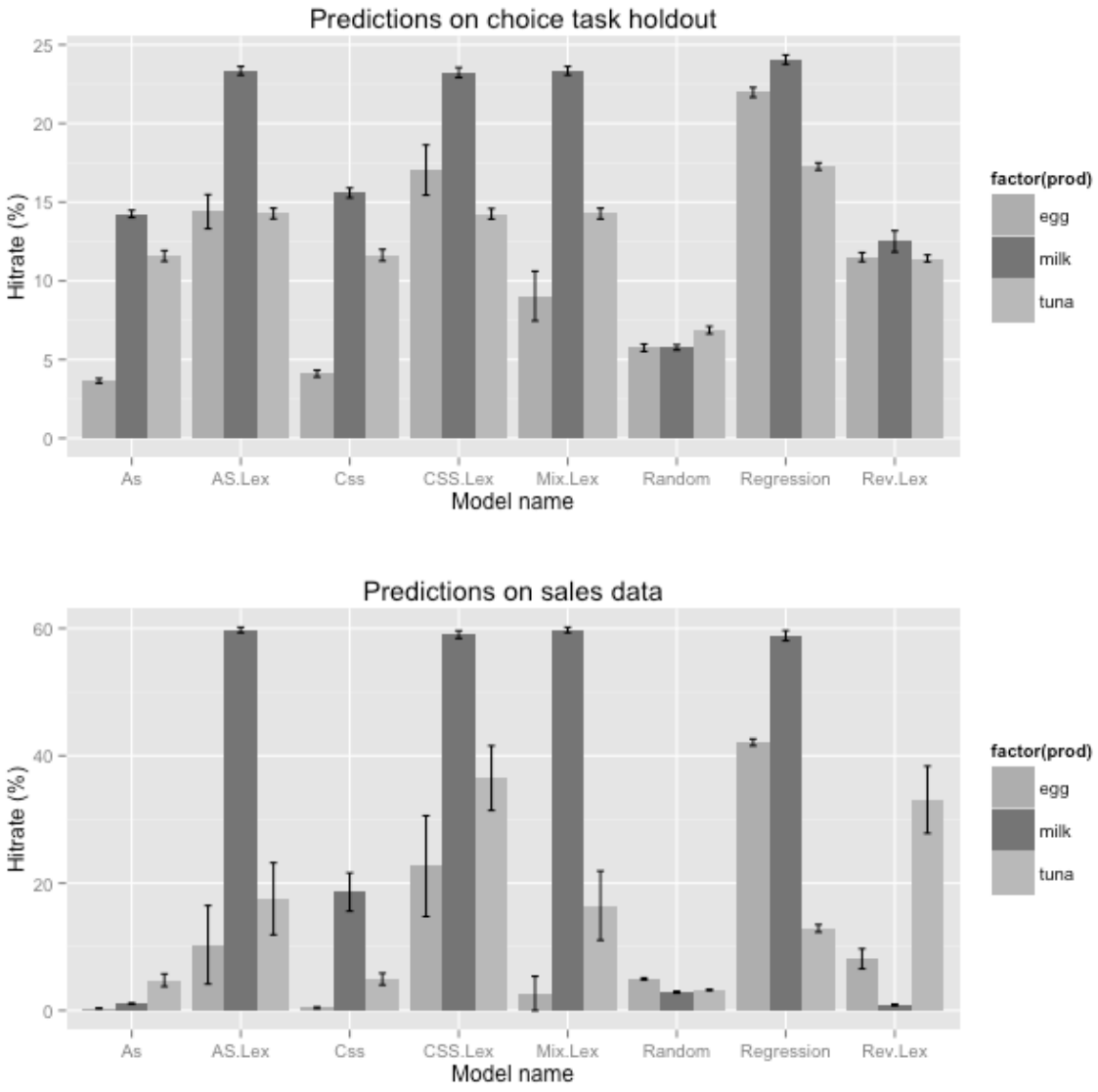


Figure 19 Summary of predictive models' hit-rates, parameterized on 50% of the respondents at each run (30 runs).

Appendix 2. Full wording of the survey for Chapter 5.

Below is the full translation of the questions in the survey to English

Please enter your age:

What is your overall opinion of the following countries:
From 0: very bad to 10 very good (with 5 being neutral)
USA
Russia
Britain
India
Pakistan
Saudi Arabia
China

Please watch the following short video:

Have you seen this video before?

On a scale from 0 to 10, how much did you like the video?
from 0: did not like the video at all, to 10: like the video a lot

On a scale from 0 to 10, how much did you understand the video?
from 0: did not understand the video at all, to 10: completely understood the video

On a scale from 0 to 10, how much did you trust the message the video?
from 0: did not trust the video at all, to 10: completely trust the video

Polio is a highly infectious disease caused by a virus and can lead to irreversible paralysis. There is no general cure for polio, but it can be prevented through immunisation. That is why many countries sponsor such programs all around the world with the aim of eradicating polio.
In general, are you for or against such immunization programs?

Next, we are going to show you six pairs of polio educational initiatives, and ask you which one you would hypothetically support. Please note that each choice is independent from the last. There is no right or wrong answer here, we are just interested in your opinion.

Consider the following two polio educational programs:
Please choose which one you would prefer (if both options are unappealing, please choose the one that you dislike the least).
Plan A: A program sponsored by [country], valued at Rs [amount] lakh. Running for 5 years starting in [month]
Plan B: A program sponsored by [country], valued at Rs [amount] lakh. Running for 5 years starting in [month]
Which of the two programs would you support?

Suppose you get a car as a reward, and four cars of the same size are placed in front of you. Which country's car would you like to have?
Japan
USA
Britain
Russia

What is your overall opinion of the following countries:
From 0: very bad to 10 very good (with 5 being neutral)
USA
Russia
Britain
India
Pakistan
Saudi Arabia
China

What is your overall trust in the following countries:
From 0: no trust, to 10: a lot of trust (with 5 being neutral)
Russia
Saudi Arabia
USA
India

Pakistan
Britain
China

In making international policy decisions, to what extent do you think the following countries take into account interests of countries like Pakistan?
Russia
Britain
USA
China

Overall, do you think the following countries are more of a partner of Pakistan, more an enemy of Pakistan, or neither?
Russia
Britain
USA
China

Please tell me if you think each of the following countries is having a mainly positive or mainly negative influence in the world?
Russia
Britain
USA
China

A number of countries have sent aid to Pakistan over the years. Please check which country you would be happy for us anonymously thank for their assistance to Pakistan.
Check all that apply, remember that this is completely anonymous.
China
Britain
USA
Saudi Arabia
None

Finally, we will ask you a few short questions to get to know you better
What is your gender?

What year were you born?

What is the highest level of education you have completed?

What is your marital status?

How many children do you have?

Ideally, if you had the opportunity, would you like to move permanently to another country, or would you prefer to continue living in this country?

To which country would you like to move?

What is your marital status?

How many children do you have?

In your day to day life, are you able to see videos on the internet (on websites such as Youtube)?

Do you watch or read news in English?

Do you watch or read news from the CNN or BBC?

Which of the following news sources do you trust?

This is the end of the survey. Thank you very much for your participation.

References:

Aaker, D. A., & Stayman, D. M. (1990). Measuring audience perceptions of commercials and relating them to ad impact. *Journal of Advertising Research*, *30*(4), 7–18.

Abernethy, A. M., & Franke, G. R. (1996). The information content of advertising: A meta-analyis. *Journal of Advertising*, *25*(2), 1–17.

Abraham, A., & Patro, S. (2014). "Country-of-Origin" Effect and Consumer Decision-making. *Management and Labour Studies*, *39*(3), 309–318.

Adamowicz, W., Boxall, P., Williams, M., & Louviere, J. (1998). Stated preference approaches for measuring passive use values: Choice experiments and contingent valuation. *American Journal of Agricultural Economics*, *80*(1), 64–75.

Agarwal, M. K., & Green, P. E. (1991). Adaptive Conjoint Analysis Versus Self Explicated Models: Some Empirical Results. *International Journal of Research in Marketing*, *8*(2), 141–146.

Akaah, P., & Korgaonkar, P. K. (1983). An Empirical Comparison of the Predictive Validity of Self-Explicated, Huber-Hybrid, Traditional Conjoint, and Hybrid Conjoint Models. *Journal of Marketing Research*, *20*(2), 187–197.

Arcuri, L., Castelli, L., Galdi, S., Zogmaister, C., & Amadori, A. (2008). Predicting the Vote : Implicit Attitudes as Predictors of the Future Behavior of Decided and Undecided Voters. *Political Psychology*, *29*(3), 369–387.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106.

Ariely, D., Loewenstein, G., & Prelec, D. (2006). Tom Sawyer and the construction of value. *Journal of Economic Behavior & Organization*, *60*(1), 1–10.

Atkin, C. K., Bowen, L., Nayman, O. B., & Sheinkopf, K. G. (1973). Quality versus quantity in televised political ads. *Public Opinion Quarterly*, *37*(2), 209–224.

Atkin, C. K., & Heald, G. (1976). Effects of political advertising. *Public Opinion Quarterly*, *40*(2), 216–228.

Auger, P., & Devinney, T. M. (2007). Do what consumers say matter? The misalignment of preferences with unconstrained ethical intentions. *Journal of Business Ethics*, *76*(4),

361–383.

Bar-Hillel, M. (2015). Position Effects in Choice From Simultaneous Displays: A Conundrum Solved. *Perspectives on Psychological Science*, *10*(4), 419–433.

Beath, A., Christia, F., & Enikolopov, R. (2011). *Winning Hearts and Minds through Development Aid: Evidence from a Field Experiment in Afghanistan Evidence from a Field Experiment in Afghanistan* (MIT Political Science Department Research Paper No. 2011–14). *Working Paper*.

Becker, N., Tavor, T., Friedler, L., & Bar (Kutiel), P. (2016). Consumers' Preferences Toward Organic Tomatoes: A Combined Two-Phase Revealed-Stated Approach. *Journal of International Food & Agribusiness Marketing*, *28*(1), 1–17.

Bettman, J. R. (1979). *An information processing theory of consumer choice*. Reading, Mass: Addison-Wesley Pub. Co.

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive Consumer Choice Processes. *Journal of Consumer Research*, *25*(3), 187–217.

Bettman, J. R., Luce, M. F., & Payne, J. W. (2008). Preference construction and preference stability: Putting the pillow to rest. *Journal of Consumer Psychology*, *18*(3), 170–174.

Bettman, J. R., & Park, C. W. (1980). Effects of Prior Knowledge and Experience and Phase of the Choice Process on Consumer Decision Processes: A Protocol Analysis. *Journal of Consumer Research*, *7*(3), 234.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841.

Boulstridge, E., & Carrigan, M. (2000). Do consumers really care about corporate responsibility? Highlighting the attitude-behaviour gap. *Journal of Communication Management*, *4*(4), 355–363.

Brinol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, *20*(October), 49–96.

Brunette, W., Sundt, M., Dell, N., Chaudhri, R., Breit, N., & Borriello, G. (2013). Open Data Kit 2.0: Expanding and Refining Information Services for Developing Regions. In *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications* (p. 10:1-10:6). New York, NY: ACM.

Bursztyn, L., Callen, M., Ferman, B., Gulzar, S., Hasanain, A., & Yuchtman, N. (2014). *Identifying Ideology: Experimental Evidence on Anti-Americanism in Pakistan* (Working Paper Series No. 20153).

Cantor, J. R., Zillmann, D., & Bryant, J. (1975). Enhancement of experienced sexual arousal in response to erotic stimuli through misattribution of unrelated residual excitation. *Journal of Personality and Social Psychology*, *32*(1), 69–75.

Carrigan, M., & Attalla, A. (2001). The Myth of the Ethical Consumer – Do Ethics Matter in Purchase Behavior? Journal of Consumer Marketing. *Journal of Consumer Marketing*, *18*(7), 560–578.

Carrington, M. J., Neville, B. A., & Whitwell, G. J. (2010). Why ethical consumers don't walk their talk: Towards a framework for understanding the gap between the ethical purchase intentions and actual buying behaviour of ethically minded consumers. *Journal of Business Ethics*, *97*(1), 139–158.

Carruthers, P. (2009). How we know our own minds: the relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, *32*(2), 121-38-82.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Cheek, N. N., Coe-odess, S., & Schwartz, B. (2015). What have I just done ? Anchoring, self-knowledge, and judgments of recent behavior. *Judgment and Decision Making*, *10*(1), 76–85.

Chrzan, K., & Golovashkina, N. (2006). An empirical test of six stated importance measures. *International Journal of Market Research*, *48*(6), 717–741.

Crane, A., & Matten, D. (2005). *Business ethics: Managing corporate citizenship and sustainability in the age of globalization*. Oxford University Press.

Croissant, Y. (2012). Estimation of multinomial logit models in R: The mlogit Packages. *R Package Version 0.2-2 URL: Http://cran.r-Project.org/web/packages/mlogit/vignettes/mlogit.pdf*.

Cummings, R. G., & Taylor, L. O. (1999). Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *American Economic Review*, *89*(3), 649–665.

Currim, I. S., Weinberg, C. B., & Wittink, D. R. (1981). Design of Subscription Programs

for a Performing Arts Series. *Journal of Consumer Research*, *8*(1), 67.

Dasgupta, N., & Greenwald, A. G. (2001). On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals. *Journal of Personality and Social Psychology*, *81*(5), 800–814.

De Bekker-Grob, E. W., Ryan, M., & Gerard, K. (2012). Discrete choice experiments in health economics: A review of the literature. *Health Economics*, *21*(2), 145–172.

De Pelsmacker, P., Driesen, L., & Rayp, G. (2005). Do consumers care about ethics? Willingness to pay for fair-trader coffee. *The Journal of Consumers Affairs*, *39*(2), 363–386.

Delavande, A., & Zafar, B. (2015). Information and Anti-American Attitudes. *Federal Reserve Bank of New York Staff Reports, No. 558*.

DellaVigna, S., & Gentzkow, M. (2009). *Persuasion: Empirical Evidence* (NBER Working Paper No. 15298). Cambridge, MA.

DellaVigna, S., & Kaplan, E. (2007). The Fox News Effect: Media Bias and Voting. *The Quarterly Journal of Economics*, *122*(3), 1187–1234.

DeMaio, T. (1984). Social desirability and survey measurement: A review. In C. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (pp. 257–282). New York: Russell Sage Foundation.

DeMarzo, P. M., Vayanos, D., & Zwiebel, J. (2003). Persuasion Bias, Social Influence, and Unidimensional Opinions. *The Quarterly Journal of Economics*, *118*(3), 909–968.

Dhar, R., & Nowlis, S. M. (1999). The Effect of Time Pressure on Consumer Choice Deferral. *Journal of Consumer Research*, *25*(4), 369–384.

Dickson, P. R., & Sawyer, A. G. (1990). The Price Knowledge and Search of Supermarket Shoppers. *Journal of Marketing*, *54*(3), 42–53.

Dieckmann, A., Dippold, K., & Dietrich, H. (2009). Compensatory versus noncompensatory models for predicting consumer preferences. *Judgment and Decision Making*, *4*(3), 200–213.

Ding, M. (2007). An Incentive-Aligned Mechanism for Conjoint Analysis. *Journal of Marketing Research*, *44*(2), 214–223.

Ding, M., Grewal, R., & Liechty, J. (2005). Incentive-Aligned Conjoint Analysis. *Journal of*

Marketing Research.

Ding, M., Hauser, J. R., Dong, S., Dzyabura, D., Yang, Z., Su, C., & Gaskin, S. P. (2011). Unstructured Direct Elicitation of Decision Rules. *Journal of Marketing Research*, *48*(1), 116–127.

Ditonto, T. M., Lau, R. R., & Sears, D. O. (2013). AMPing racial attitudes: Comparing the power of explicit and implicit racism measures in 2008. *Political Psychology*, *34*(4), 487–510.

Ducoffe, R. H. (1995). How Consumers Assess the Value of Advertising. *Journal of Current Issues & Research in Advertising*, *17*(1), 1–18.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment implications for health, education, and the workplace. *Psychological Science in the Public Interest, Supplement*, *5*(3), 69–106.

Dutton, D. G., & Aron, A. P. (1974). Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology*, *30*(4), 510–517.

Eggers, F., & Sattler, H. (2011). Preference Measurement with Conjoint Analysis: Overview of State-of-the-art Approaches and Recent Developments. *GfK Marketing Intelligence Review*, *3*(1), 36–47.

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance*, *6*(1), 1–27.

Elrod, T., Louviere, J. J., & Davey, K. S. (1992). An empirical comparison of ratings-based and choice-based conjoint models. *Journal of Marketing Research*, *29*(3), 368–377.

Essoussi, L. H., & Zahaf, M. (2009). Exploring the decision-making process of Canadian organic food consumers: Motivations and trust issues. *Qualitative Market Research: An International Journal*, *12*(4), 443–459.

Fazio, R. H., & Olson, M. a. (2003). Implicit measures in social cognition. research: their meaning and use. *Annual Review of Psychology*, *54*, 297–327.

Fischhoff, B., Slovic, P., Lichtenstein, S., Bell, D. E., Raiffa, H., & Tversky, A. (1988). Knowing what you want: Measuring labile values. In T. S. Wallsten (Ed.), *Cognitive Processes in Choice and Decision Behavior* (pp. 117–142). Erlbaum.

Frederick, S., Lee, L., & Baskin, E. (2014). The Limits of Attraction. *Journal of Marketing Research*, *51*(4), 487–507.

Friese, M., Wänke, M., & Plessner, H. (2006). Implicit Consumer Preferences and Their Influence on Product Choice. *Psychology & Marketing*, *23*(9), 727–740.

Gadema, Z., & Oglethorpe, D. (2011). The use and usefulness of carbon labelling food: A policy perspective from a survey of UK supermarket shoppers. *Food Policy*, *36*(6), 815–822.

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. *Science*, *321*(5892), 1100–2.

Gallup Pakistan. (2016). *Questions Related United State (1979 – 2016)*.

Gawronski, B. (2002). What does the Implicit Association Test measure? A test of the convergent and discriminant validity of prejudice-related IATs. *Experimental Psychology*, *49*(3), 171–180.

Gawronski, B., LeBel, E., & Peters, K. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science*, *2*(2), 181–193.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gilbride, T. J., & Allenby, G. M. (2004). A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules. *Marketing Science*, *23*(3), 391–406.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*(1), 1–14.

Green, P. E., Goldberg, S. M., & Wiley, J. B. (1982). A Cross-Validation Test of Hybrid Conjoint Models. *Advances in Consumer Research*, *10*, 147–150.

Green, P. E., & Helsen, K. (1989). Cross-Validation Assessment of Alternatives to Individual-Level Conjoint Analysis: A Case Study. *Journal of Marketing Research*, *26*(3), 346–350.

Green, P. E., Krieger, A. M., & Agarwal, M. K. (1993). A cross validation test of four

models for quantifying multiattribute preferences. *Marketing Letters*, *4*(4), 369–380.

Green, P. E., Krieger, A. M., & Bansal, P. (1988). Completely Unacceptable Levels in Conjoint Analysis: A Cautionary Note. *Journal of Marketing Research*, *25*(3), 293–300.

Green, P. E., & Rao, V. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, *8*, 355–363.

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

Green, P. E., & Srinivasan, V. (1990). Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice. *Journal of Marketing*, *54*(4), 3.

Green, P. E., Wind, Y., & Carmone, F. J. (1972). Subjective evaluation models and conjoint measurement. *Behavioral Science*, *17*(3), 288–299.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring Individual Differences in Implicit Cognition : The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1–20.

Harrison, R., Newholm, T., & Shaw, D. (Eds.). (2005). *The Ethical Consumer*. London: Sage Publications.

Hauser, J. R. (2014). Consideration-set heuristics. *Journal of Business Research*, *67*(8), 1688–1699.

Hauser, J. R., Dong, S., & Ding, M. (2014). Self-reflection and articulated consumer preferences. *Journal of Product Innovation Management*, *31*(1), 17–32.

Hensel-Börner, S., & Sattler, H. (1999). *Validität der Customized Computerized Conjoint*

Analysis (CCC).

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

Huber, G. P. (1974). Multi-Attribute Utility Models: A Review of Field and Field-Like Studies. *Management Science*, *20*(10), 1393–1402.

Huber, G. P., Daneshgar, R., & Ford, D. L. (1971). An empirical comparison of five utility models for predicting job preferences. *Organizational Behavior and Human Performance*, *6*(3), 267–282.

Huber, J., Payne, J. W., & Puto, C. (1982). Alternatives : Violations of Regularity and the Similarity Hypothesis. *The Journal of Consumer Research*, *9*(1), 90–98.

Huber, J., Wittink, D. R., Fiedler, J. A., & Miller, R. (1993). The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice. *Journal of Marketing Research*, *30*(1), 105–114.

Hudson, D., Gallardo, R. K., & Hanson, T. R. (2012). A Comparison of Choice Experiments and Actual Grocery Store Behavior : An Empirical Application to Seafood Products. *Journal of Agricultural and Applied Economics*, *44*(1), 49–62.

Jain, A. K., Acito, F., Malhotra, N. K., & Mahajan, V. (1979). A Comparison of the Internal Validity of Alternative Parameter Estimation Methods in Decompositional Multiattribute Preference Models. *Journal of Marketing Research*, *16*(3), 313–322.

Johansson, P., Hall, L., & Sikström, S. (2008). From change blindness to choice blindness. *Psychologia*, *51*(2), 142–155.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–119.

Johnston, R. J., Wessells, C. R., Donath, H., & Asche, F. (2001). Measuring consumer preferences for ecolabeled seafood: an international comparison. *Journal of Agricultural and Resource Economics*, *26*(1), 20–39.

Joshi, Y., & Rahman, Z. (2015). *Factors Affecting Green Purchase Behaviour and Future Research Directions*. *International Strategic Management Review* (Vol. 3). Holy Spirit University of Kaslik.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*,

39(4), 341–350.

Kivetz, R., Netzer, O., & Schrift, R. (2008). The synthesis of preference: Bridging behavioral decision research and marketing science. *Journal of Consumer Psychology*, *18*(3), 179–186.

Kohli, R., & Jedidi, K. (2007). Representation and Inference of Lexicographic Preference Models and Their Variants. *Marketing Science*, *26*(3), 380–399.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the Accuracy of Personality Judgments by the Self and Knowledgeable Others. *Journal of Personality*, *64*(2), 311–337.

Kozuch, B., & Nichols, S. (2011). Awareness of Unawareness Folk Psychology and Introspective Transparency. *Journal of Consciousness Studies*, *18*(11–12).

Krueger, A. B., & Malecková, J. (2009). Attitudes and action: public opinion and the occurrence of international terrorism. *Science (New York, N.Y.)*, *325*(5947), 1534–1536.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134.

Leach, S., & Weick, M. (2017). Can People Judge the Veracity of Their Intuitions? *Social Psychological and Personality Science*, 1–10.

Leigh, T., MacKay, D., & Summers, J. (1984). Reliability and validity of conjoint analysis and self-explicated weights: A comparison. *Journal of Marketing Research*, *21*(November), 456–463.

Levitt, S. D., & List, J. A. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World ? *The Journal of Economic Perspectives*, *21*(2), 153–174.

Lichtenstein, S., & Slovic, P. (2006). *The Construction of Preference* (Cambridge).

List, J. A., & Gallet, C. A. (2001). What experimental protocol influence disparities between actual and hypothetical stated values? *Environmental and Resource Economics*, *20*(3), 241–254.

Little, J., & Berrens, R. (2003). Explaining disparities between actual and hypothetical stated

values: Further investigation using meta-analysis. *Economics Bulletin*, *3*(1).

Lodge, M. (1981). *Magnitude Scaling: Quantitative Measurement of Opinions*. *Sage University Paper Series on Quantitative Applications in the Social Sciences*. Beverly Hills: Sage Publications.

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods: analysis and applications*. *Analysis and Applications* (Vol. 12). Cambridge University Press.

Louviere, J. J., & Islam, T. (2008). A comparison of importance weights and willingness-to-pay measures derived from choice-based conjoint, constant sum scales and best-worst scaling. *Journal of Business Research*, *61*, 903–911.

Luce, R. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Luchini, S., & Watson, V. (2014). Are choice experiments reliable? Evidence from the lab. *Economics Letters*, *124*(1), 9–13.

Lusk, J. L. (2003). Effects of Cheap Talk on Consumer Willingness to pay for Golden Rice. *American Journal of Agricultural Economics*, *85*(4), 840–856.

Lusk, J. L., & Briggeman, B. C. (2009). Food values. *American Journal of Agricultural Economics*, *91*(1), 184–196.

Lusk, J. L., Pruitt, J. R., & Norwood, B. (2006). External validity of a framed field experiment. *Economics Letters*, *93*(2), 285–290.

Lusk, J. L., & Schroeder, T. C. (2004). Are choice experiments incentive compatible? A test with quality differentiated beef steaks. *American Journal of Agricultural Economics*, *86*(2), 467–482.

Maier, M., Maier, J., Baumert, A., Jahn, N., Krause, S., & Adam, S. (2015). Measuring citizens' implicit and explicit attitudes towards the European Union. *European Union Politics*, *16*(3), 369–385.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? A case study in ecological rationality. In *Simple heuristics that make us smart Gigerenzer Gerd* (pp. 119–140).

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In *Prejudice, Discrimination, and Racism* (pp. 91–125).

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers*

in Econometrics (pp. 105–142).

McFadden, D. (1986). The choice theory approach to market research. *Marketing Science*, *5*(4), 275–297.

Menon, G., Raghubir, P., & Schwarz, N. (1995). Behavioral Frequency Judgments: An Accessibility-Diagnosticity Framework. *Journal of Consumer Research*, *22*(September).

Mesías, F. J., Martínez-Carrasco, F., Martínez, J. M., & Gaspar, P. (2011). Functional and organic eggs as an alternative to conventional production: A conjoint analysis of consumers' preferences. *Journal of the Science of Food and Agriculture*, *91*(3), 532–538.

Meyer, J. H. (1999). *Measuring Customer Satisfaction: Hot Buttons and Other Measurement Issues*. Chicago: American Marketing Association.

Moser, A. K. (2016). Consumers' purchasing decisions regarding environmentally friendly products: An empirical analysis of German consumers. *Journal of Retailing and Consumer Services*, *31*, 389–397.

Moser, R., Raffaelli, R., & Notaro, S. (2014). Testing hypothetical bias with a real choice experiment using respondents' own money. *European Review of Agricultural Economics*, *41*(1), 25–46.

Mueller, S., Lockshin, L., & Louviere, J. J. (2009). What you see may not be what you get: Asking consumers what matters may not reflect what they choose. *Marketing Letters*, *21*(4), 335–350.

Murphy, J. J., Allen, P. G., Stevens, T. H., & Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, *30*(3), 313–325.

Næs, T., Brockhoff, P. B., & Tomic, O. (2010). *Statistics for Sensory and Consumer Science*. John Wiley & Sons.

Netzer, O., & Srinivasan, V. (2011). Adaptive self-explication of multi-attribute preferences. *Journal of Marketing Research*, *48*(1), 140–156.

Netzer, O., Toubia, O., Bradlow, E. T., Dahan, E., Evgeniou, T., Feinberg, F. M., … Rao, V. R. (2008). Beyond conjoint analysis: Advances in preference measurement. *Marketing Letters*, *19*(3–4), 337–354.

Nisbett, R. E., & Schachter, S. (1966). Cognitive manipulation of pain. *Journal of Experimental Social Psychology*, *2*(3), 227–236.

Nisbett, R. E., & Wilson, T. D. (1977a). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Nisbett, R. E., & Wilson, T. D. (1977b). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250–256.

Nosek, B. a. (2007). Implicit – Explicit Relations. *Current Directions in Psychological Science*, *16*(2), 65–69.

Padel, S., & Foster, C. (2005). Exploring the gap between attitudes and behaviour: Understanding why consumers buy or do not buy organic food. *British Food Journal*, *107*(8), 606–625.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277–293.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. *The Adaptive Decision Maker*. Cambridge: Cambridge University Press.

Payne, J. W., Bettman, J. R., & Schkade, D. A. (1999). Measuring constructed preferences: Towards a building code. *Journal of Risk and Uncertainty*, *19*(1–3), 243–270.

Petty, R. E., & Wegener, D. T. (1998). Attitude change: Multiple roles for persuasion variables. *The Handbook of Social Psychology*, (June 2016), 323–390.

Pew Global Attitudes Project. (2013). *America's Global Image Remains More Positive than China's*.

Punj, G. N., & Staelin, R. (1978). The Choice Process for Graduate Business Schools. *Journal of Marketing Research*, *XV*(November), 588–599.

Resnik, A., & Stern, B. L. (1977). An Analysis of Information Content in Television Advertising. *Journal of Marketing*, *41*(1), 50.

Roberts, J. A. (1995). Profiling Levels of Socially Responsible Consumer Behavior: A Cluster Analytic Approach and Its Implications for Marketing. *Journal of Marketing Theory and Practice*, *3*(4), 97–117.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of

explanatory depth. *Cognitive Science*, *26*(5), 521–562.

Rudman, L. A., & Kilianski, S. E. (2000). Implicit and Explicit Attitudes Toward Female Authority. *Personality and Social Psychology Bulletin*, *26*(11), 1315–1328.

Sachdeva, S., Jordan, J., & Mazar, N. (2015). Green consumerism: moral motivations to a sustainable future. *Current Opinion in Psychology*, *6*, 60–65.

Sattler, H., & Hensel-Börner, S. (2007). A comparison of conjoint measurement with self-explicated approaches. In A. Gustafsson, A. Herrmann, & F. Huber (Eds.), *Conjoint Measurement* (4th ed., pp. 67–76). Berlin: Springer Berlin Heidelberg.

Schachter, S., & Singer, J. E. (1962). Cognitive, social, and psychological determinants of emotional state. *Psychological Review*, *69*, 379–399.

Scheibehenne, B., Miesler, L., & Todd, P. M. (2007). Fast and frugal food choices: Uncovering individual decision heuristics. *Appetite*, *49*(3), 578–589.

Schmidt, U., Starmer, C., & Sugden, R. (2008). Third-generation prospect theory. *Journal of Risk and Uncertainty*, *36*(3), 203–223.

Schulz, E., Speekenbrink, M., & Shanks, D. R. (2014). Predict choice: A comparison of 21 mathematical models. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2889–2894.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: experiments on question form, wording, and context*. New York: Academic Press.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist*, *54*(2), 93–105.

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly*, *49*(3), 388–395.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating Scales Numeric Values May Change the Meaning of Scale Labels. *Public Opinion Quarterly*, *55*(4), 570–582.

Schwarz, N., & Oyserman, D. (2001). Asking Questions About Behavior: Cognition, Communication, and Questionnair Construction. *American Journal of Evaluation*, *22*(2), 127–160.

Schwarz, N., & Scheuring, B. (1992). Selbstberichtete Verhaltens- und Symptomhäufigkeiten: Was Befragte aus Antwortvorgaben des Fragebogens lernen. *Zeitschrift Für Klinische Psychologie*, *21*(2), 197–208.

Shaw, D., Grehan, E., Shiu, E., Hassan, L., & Thomson, J. (2005). An exploration of values in ethical consumer decision making. *Journal of Consumer Behaviour*, *4*(3), 185–200.

Shaw, J. I., Bergen, J. E., Brown, C. A., & Gallagher, M. E. (2000). Centrality preferences in choices among similar options. *Journal of General Psychology*, *127*(2), 157–164.

Sheeran, P. (2002). Intention—Behavior Relations: A Conceptual and Empirical Review. *American Journal of Public Health*, *12*(1), 1–36.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074.

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research*, *16*(2), 158.

Simonson, I. (2008a). Regarding inherent preferences. *Journal of Consumer Psychology*, *18*(3), 191–196.

Simonson, I. (2008b). Will I like a "medium" pillow? Another look at constructed and inherent preferences. *Journal of Consumer Psychology*, *18*(3), 155–169.

Simonson, I. (2014). Vices and Virtues of Misguided Replications: The Case of Asymmetric Dominance. *Journal of Marketing Research*, *51*(4), 514–519.

Sirieix, L., Kledal, P. R., & Sulitang, T. (2011). Organic food consumers' trade-offs between local or imported, conventional or organic products: a qualitative study in Shanghai. *International Journal of Consumer Studies*, *35*(6), 670–678.

Smith, E. R. (2008). Embodied constraints as elements in attitude construction. *Journal of Consumer Psychology*, *18*(3), 187–190.

Srinivasan, V. (1988). A Conjunctive-Compensatory Approach To the Self-Explication of Multiattributed Preferences. *Decision Sciences*, *19*(2), 295–305.

Srinivasan, V., & Park, C. S. (1997). Surprising Robustness of the Self-Explicated Approach to Consumer Preference Structure Measurement. *Journal of Marketing Research*,

34(2), 286–291.

Srivastava, J., Connolly, T., & Beach, L. R. (1995). Do Ranks Suffice? A Comparison of Alternative Weighting Approaches in Value Elicitation. *Organizational Behavior and Human Decision Processes*.

Stephenson, W. (1953). *The study of behavior; Q-technique and its methodology.* Chicago, IL: University of Chicago Press.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29.

Tourangeau, R., & Smith, T. W. (1996). Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context. *Public Opinion Quarterly*, *60*(2), 275.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31–48.

Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, *79*(4), 281–299.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1986). Rational Choice and the Framing of Decisions. *The Journal of Business*, *54*(4), S251–S278.

USAID. (2015). *U.S. Overseas Loans and Grants: Obligations and Loan Authorizations, July 1, 1945–September 30, 2015*.

Valenzuela, A., & Raghubir, P. (2009). Position-based Beliefs: The Center-Stage Effect. *Journal of Consumer Psychology*, *19*(2), 185–196.

Van Ittersum, K., Pennings, J. M. E., Wansink, B., & van Trijp, H. C. M. (2007). The validity of attribute-importance measurement: A review. *Journal of Business Research*, *60*(11), 1177–1190.

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300.

Vazire, S., & Carlson, E. N. (2011). Others Sometimes Know Us Better Than We Know Ourselves. *Current Directions in Psychological Science*, *20*(2), 104–108.

Verlegh, P. W. J., Schifferstein, H. N. J., & Wittink, D. R. (2002). Range and number-of-levels effects in derived and stated measures of attribute importance. *Marketing Letters*, *13*(1), 41–52.

Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. a. (2011). Does the brain calculate value? *Trends in Cognitive Sciences*, *15*(11), 546–54.

Voeth, M., Herbst, U., & Liess, F. (2013). We know exactly what you want: the development of a completely individualised conjoint analysis. *International Journal of Market Research*, *55*(3), 437.

Von Neumann, J., & Morgenstern, O. (1944). Theory of Games and Economic Behavior. *Princeton University Press*, 625.

Voors, M., Turley, T., Kontoleon, A., Bulte, E., & List, J. A. (2012). Exploring whether behavior in context-free experiments is predictive of behavior in the field: Evidence from lab and field experiments in rural Sierra Leone. *Economics Letters*, *114*(3), 308–311.

Westenberg, M. R. M., & Koele, P. (1994). Multi-attribute evaluation processes: Methodological and conceptual issues. *Acta Psychologica*, *87*, 65–84.

Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. *Advances in Experimental Social Psychology, Vol. 22*.

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: its limits, value, and potential for improvement. *Annual Review of Psychology*, *55*, 493–518.

Wilson, T. D., Lisle, D. J., Schooler, J. W., Hodges, S. D., Klaaren, K. J., & Lafleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin*, *19*(3), 331–339.

Wilson, T. D., & Nisbett, R. E. (1978). The Accuracy of Verbal Reports About the Effects of Stimuli on Evaluations and Behavior. *Social Psychology*, *41*(2), 118–131.

Wilson, T. D., & Schooler, J. W. (1991). Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decisions. *Journal of Personality and Social Psychology*, *60*(2), 181–192.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology*, *81*(5), 815–827.

Wittink, D. R., Vriens, M., & Burhenne, W. (1994). Commercial use of conjoint analysis in Europe: Results and critical reflections. *International Journal of Research in Marketing*, *11*(1), 41–52.

Wolfe, M. B., & Williams, T. J. (2017). Poor Metacognitive Awareness of Belief Change. *The Quarterly Journal of Experimental Psychology*, *9*, 1–45.

Wright, P., & Kriewall, M. A. (1980). State-of-mind effects on the accuracy with which utility functions predict marketplace choice. *Journal of Marketing Research*, *17*(August), 277–293.

Yee, M., Dahan, E., Hauser, J. R., & Orlin, J. (2007). Greedoid-Based Noncompensatory Inference. *Marketing Science*, *26*(4), 532–549.

Zanoli, R., & Naspetti, S. (2002). Consumer motivations in the purchase of organic food: A means-end approach. *British Food Journal*, *104*(8), 643–653.

Zepeda, L., & Deal, D. (2009). Organic and local food consumer behaviour: Alphabet Theory. *International Journal of Consumer Studies*, *33*(6), 697–705.