

A Choice Matching Approach for Discrete Choice Analysis: An Experimental Investigation in the Lab

Simone Cerroni, Associate Professor, Department of Economics and Management and C3A, University of Trento (Italy), Email: simone.cerroni@unitn.it, Full postal address: Via Inama 5, Trento, 38122, Italy (Corresponding author)

Daniel Derbyshire, Research Fellow, European Centre for Environment and Human Health, University of Exeter, Email: D.W.Derbyshire@exeter.ac.uk, Full postal address: Knowledge Spa, Royal Cornwall Hospital Truro, Cornwall, TR1 3HD, United Kingdom.

W. George Hutchinson, Professor, Gibson Institute and Institute for Global Food Security, Queen's University Belfast, Email: g.hutchinson@qub.ac.uk, Full postal address: 19 Chlorine Gardens, Belfast BT9 5DL, United Kingdom.

Rodolfo M. Nayga, Jr., Professor and Department Head, Department of Agricultural Economics, Texas A&M University, Email: Rnayga@tamu.edu, Full postal address: Agriculture & Life Science Bldg, Suite 309 2424, 600 John Kimbrough Blvd, College Station, TX 77843, United States

Abstract

This paper is the first empirical application of the choice matching (CM) method in discrete choice experiments (DCE). An artefactual field experiment was conducted to test whether the CM applied to a DCE survey improves the validity and reliability of estimated preferences with respect to standard hypothetical DCE. Two experimental treatments were developed. In the first, subjects were exposed to the CM-based DCE; in the second, to a standard hypothetical DCE survey. Results suggest that while the CM-based DCE does not improve validity, it can increase the reliability of estimated preferences.

Keywords: discrete choice experiment, choice matching, honest responses, hypothetical bias, validity, reliability

JEL: C83, C91, D12

Appendix materials can be accessed online at:

<https://uwpress.wisc.edu/journals/pdfs/LE-99-1-Cerroni-appA.pdf>

<https://uwpress.wisc.edu/journals/pdfs/LE-99-1-Cerroni-appB.pdf>

<https://uwpress.wisc.edu/journals/pdfs/LE-99-1-Cerroni-appC.pdf>

<https://uwpress.wisc.edu/journals/pdfs/LE-99-1-Cerroni-appD.pdf>

<https://uwpress.wisc.edu/journals/pdfs/LE-99-1-Cerroni-appE.pdf>

1. Introduction

Stated preference (SP) methods are widely used in many branches of applied economics, ranging from environmental to agri-food economics, and from health to transportation economics. SP can evaluate consumer preferences and willingness to pay (WTP) for innovative products that are not offered on the market yet and evaluate *ex ante* the economic value of welfare benefits generated by public policies that are not yet implemented. Such evaluations can be very useful for businesses and policy makers and therefore need to be accurately estimated.

The accuracy of SP's results, however, has often been questioned in the literature (e.g., Harrison and Rutström 2008, Harrison 2014). This paper contributes to this literature by providing the first empirical application of the choice matching (CMa) method to discrete choice experiments (DCE) that is arguably the most used SP technique. The CMa method was recently developed by Cvitanić et al. (2019) to elicit honest responses using any type of discrete choice question and could improve the accuracy of preferences elicited using DCE surveys by bridging the gap between stated and revealed preference methods. The CMa can be considered a refinement of Prelec's Bayesian Truth Serum (BTS) (2004). In this paper, an artefactual field experiment¹ was used to compare the accuracy which is measured in terms of validity and reliability between the CMa applied to a DCE survey and a standard hypothetical DCE survey.

The accuracy of SP methods is often criticized for two major reasons. First, rational choice theory which is the theoretical foundation of these methods has been seriously challenged by empirical evidence from psychology and behavioral economics (e.g., Camerer 1995; 1999). Second, the hypothetical nature of the setting where respondents are asked to make decisions undermines the incentive compatibility of most SP approaches. In such hypothetical settings, truthful responses to the survey questions may no longer be the optimal strategy for respondents (Carson and Groves 2007) and could generate hypothetical bias (HB), which is the discrepancy between behavior observed in hypothetical and real choice settings (Harrison, Harstad, and Rutström 2004). HB often

leads to an overestimation of WTP with respect to market settings where real transactions occur (List and Gallet 2001; Murphy et al. 2005; Penn and Hu 2018).

Experimental methods are often used to further explore these issues and can contribute to SP literature in at least two ways (Harrison and Rutström 2008, Harrison 2014). First, carefully designed experiments conducted in controlled environments can be used to test whether rational choice theory is supported when people face valuation exercises and identify the conditions that facilitate the satisfaction of rational choice theory's assumptions (Shogren 2005; 2006). Second, experimental methods can be used to assess the extent of HB, provide best practices to mitigate HB and test the efficacy of approaches developed to minimize HB (Harrison and Rutström 2008, Harrison 2014).

A few caveats, however, must be considered when using economic experiments for the latter purpose. While experimental procedures used to elicit values and preferences are theoretically incentive compatible mechanisms, they may not be fully demand revealing in practice (e.g., Cerroni et al. 2019). This raises the issue of whether economic experiments are able to elicit respondents' true values and preferences (e.g., Harrison, Harstad, and Rutström 2004). Furthermore, experimental methods cannot be easily implemented in many non-market valuation contexts (e.g., environmental and health economics) because the goods and services under valuation cannot be exchanged for money at the end of the experiment, either because these are not available or these are too costly. Hence, experimental methods can be rarely used to elicit values for public goods that are often the main focus of SP applications (e.g., Adamowicz 2004; Shogren 2005). Finally, while controlled laboratory experiments using standard subject pools (i.e., students) are characterized by a high degree of internal validity, their ability to draw generalizable conclusions that are meaningful to explain behavior in "real-life" situations is often questioned (e.g., Gneezy and Imas 2017).² To overcome this limitation, some experimental research has tried to move away from the lab to the field and add context to experimental designs (e.g., Harrison and List 2004). However, this move comes at a cost: the introduction of possible confounding factors and lower degree of internal

validity (e.g., Smith 1976). Economic experiments in the area of non-market valuation are always prone to the tension between laboratory control (i.e., internal validity) and natural context (i.e., external validity) (Shogren 2010).

Despite these caveats, the use of experimental methods to investigate HB in SP surveys has stimulated the development of several methods to mitigate the problem. These are categorized into *ex ante* methods that aim to reduce hypothetical bias by survey design and *ex post* methods that aim to correct potentially biased preferences using calibration approaches (see Loomis 2014 for a review). A number of *ex ante* approaches to reducing HB in DCEs exist: i) cheap talk (e.g., Cummings and Taylor 1999; Silva et al. 2011), ii) consequentiality (e.g., Vossler, Doyon, and Rondeau 2012), iii) honesty priming (e.g., De-Magistris, Gracia and Nayga Jr. 2013), iv) oaths (Jacquemet et al 2013), v) virtual reality (Fang et al. 2020), vi) indirect questioning (IQ) (e.g., Lusk and Norwood 2009a), and vii) Bayesian truth serums (BTS) (e.g., Prelec 2004). However, results on the efficacy of these methods in reducing HB are generally mixed.

Given the importance of finding new ways to reduce HB in DCEs, this paper tests whether the CMA method applied to a DCE survey improves the accuracy (in terms of validity and reliability) of elicited preferences as compared to a standard hypothetical DCE. Our empirical application focuses on consumers' preferences for a ready meal product that varies in terms of price, saturated fat, salt and whether the beef was produced with or without antibiotics. Our experimental design consists of two treatments. In the CMA treatment, subjects were exposed to the CMA method applied to a DCE survey (CMA-based DCE). In the DCE treatment, subjects were exposed to a standard hypothetical DCE survey.

The CMA method consists of two tasks. The first task (i.e., preference task) is equivalent to a standard hypothetical DCE. In the second task (i.e., belief task), respondents are asked to predict the choices that all other respondents in their session made in each choice situation presented in the first task. These predictions are elicited using an incentivized proper scoring rule (i.e., quadratic scoring rule) and respondents receive a payment depending on how accurate their predictions are. This

induces respondents to reveal their own true beliefs about other respondents' choices. The key mechanism that makes a hypothetical DCE incentive compatible under the CMA and therefore able to elicit truthful choices is the following. Respondents are informed that one choice situation will be selected at random at the end of the experiment (i.e., binding choice situation) and there is a chance that respondents' predictions in the binding choice situation will be replaced by the average predictions of the other respondents that made the same choice as them in the binding choice situation in the preference task. These predictions are in turn used to calculate the respondents' payoff from the belief elicitation task; hence respondents have strict incentives to make honest choices in the preference task. The incentive compatibility of this mechanism relies on a key assumption, impersonal updating, which requires respondents that make the same choices have similar beliefs about other respondents' choices. This key assumption is discussed in detail in Section 3 and tested empirically in Section 6. When the impersonal updating assumption is satisfied, the application of the CMA to DCE allows the elicitation of truthful choices related to any type of goods and services, including public goods that cannot be marketed. This implies that the CMA-based DCE could be potentially used in all fields of applied economics that make use of DCE.

To assess the relative validity and reliability of preferences elicited via the CMA-based DCE, random parameter logit (RPL) models (mixed logit) were estimated in willingness to pay (WTP) space. Validity and reliability are two concepts widely used in the SP literature to assess the accuracy of results (Mitchell and Carson, 1989). Reliability is related to the variance of estimated preferences and can be assessed using the standard errors of the estimated coefficients and/or error variance (i.e., scale parameter in DCE). Validity is about the truthfulness of estimated preferences. Since true values of goods under examination are often unobservable, indirect approaches must be used to assess validity and a number of criteria are available for this purpose: content validity, construct validity, convergent validity and criterion validity (Johnstone et al. 2017; Bishop and Boyle 2019; Mariel et al. 2021)

Our hypothesis is that the means of estimated marginal WTP distributions obtained via the CMA-based DCE will be lower than those obtained via a standard hypothetical DCE. This hypothesis is based on previous empirical evidence suggesting that hypothetical DCE are generally affected by HB. Regarding reliability, we expect that standard deviations of the marginal WTP distributions and standard errors of estimated coefficients via the CMA-based DCE will be lower than those obtained from the hypothetical DCE. In addition, we hypothesize that the scale parameters estimated using data from the CMA-based DCE are higher than those estimated using data collected via the hypothetical DCE. This will signal that the CMA-based DCE produces lower error variance. Our results generally suggest that while the CMA method does not necessarily improve validity, it has the potential to increase reliability of estimated preferences and WTPs.

The rest of the paper is laid out as follows. We first review methods used to reduce hypothetical bias and elicit reliable responses. Next, we present the CMA method and our application to DCEs. Then, our experimental results are presented and analysed in detail. Finally, we offer some conclusions based on the experimental results we have obtained.

2. Background

Reliability and validity in DCEs

Reliability and validity have been extensively tested in many fields of applied economics where the use of SP methods is popular such as environmental, health, energy, transportation and agri-food economics (see Table 1). The two most recent and exhaustive review studies on the reliability and validity of SP research were developed by Bishop and Boyle (2019) for the contingent valuation method (CVM) and Mariel et al. (2021) for DCEs. Johnstone et al.'s (2017) contemporary guidance for SP studies is another important source of information on the topic. Other studies have reviewed evidence on these two key concepts for specific disciplines, for example, Rakotonarivo, Schaafsma, and Hockley (2016) focuses on environmental valuation, while Janssen et al. (2017) on health economics.

In the SP literature, reliability has been interpreted in terms of consistency of values, choices and preferences. Most of this literature has explored intertemporal reliability using test-retest experiments where the same survey is conducted at a time t and replicated at a time $t+1$. A preference elicitation method is deemed to be reliable when values and preferences elicited at time t and $t+1$ are not statistically different. A large number of studies have used this approach in the DCE literature. All disciplines where the use of DCE is widespread are represented: health economics (e.g., Bryan et al. 2000; Ryan et al. 2006; Skjoldborg, Lauridsen, and Junker 2009; Price, Dupont, and Adamowicz 2017), environmental economics (e.g., Bliem, Getzner, and Rodiga-Laßnig 2012; Schaafsma et al. 2014; Matthews, Scarpa, and Marsh 2017; Brouwer, Logar, and Sheremet 2017), energy economics (e.g., Liebe, Meyerhoff, and Hartje 2012), agro-food economics (e.g., Mørkbak and Olsen 2015; Rigby, Burton, and Pluske 2016), and transportation economics (e.g., Börjesson 2014).

This approach, however, departs from Mitchell and Carson's (1989) original interpretation of reliability. Mitchell and Carson (1989) relate reliability of SP studies to the variance of elicited contingent values. Following this interpretation, reliability has been measured in several ways in the literature. Boyle and Bishop (2019) argue that reliability of the CVM can be assessed by considering the estimated standard errors of the elicited values. Specifically, larger standard errors signal lower reliability (and vice versa). Liebe, Meyerhoff, and Hartje (2012) explore the reliability of the DCE by considering the magnitude of the scale parameter which is the inverse of the error variance. Therefore, the larger the scale parameter, the lower the error variance, and the higher the reliability of elicited preferences. The idea that error variance measures reliability has been widely used in the DCE literature (e.g., Day et al. 2012; Hess, Hensher, and Daly, 2012; Campbell et al. 2015). Kealy, Montgomery, and Dovidio (1990) suggest that the variance of WTP distributions elicited via the CVM can be used as a proxy of reliability. The same criteria is used by Czajkowski, Barczak, and Budziński (2016) to test reliability of DCEs. In this paper, we utilized these

approaches to explore the reliability of the CMA-based DCE with respect to the standard hypothetical DCE.

Validity of SP surveys is related to the truthfulness of elicited values, choices, and preferences. Validity has been described in four possible ways: content validity, construct validity, convergent validity and criterion validity (Johnstone et al. 2017; Bishop and Boyle 2019; Mariel et al. 2021). Content validity focuses on the appropriateness of procedures used to design and conduct the valuation study, analyze data, and report results. Generally, content validity can be assessed in terms of adherence to best-practices highlighted in the literature (for example, Holmes, Adamowicz, and Carlsson 2017; Johnstone et al. 2017).

Construct validity focuses on prior knowledge regarding the relationship between values/preferences and other variables. This prior knowledge comes from economic theory and previous empirical studies. For example, in DCE studies, it is expected that the price coefficient is negative and statistically significant, and that lower income respondents are more sensitive to price changes in general (Mariel et al. 2021). A particular type of construct category is convergent validity which focuses on comparisons of values/preferences estimated via different value/preference-elicitation mechanisms. An example would be comparing WTPs elicited via CVM and DCE (e.g., Hanley et al. 1998; Lloyd-Smith, Zawojka, and Adamowicz 2021). The array of examples is wide and covers several disciplines: health economics (e.g., Van der Pol et al. 2008; Ryan and Watson 2009), environmental economics (e.g., Boyle et al. 2001; Caparros, Oviedo, and Campos 2008; Christie and Azevedo 2009), energy economics (e.g., McNair, Bennett, and Hensher 2011), agri-food economics (e.g., Asioli et al. 2016; Yangui et al. 2019), and transportation economics (e.g., Raffaelli et al. 2021).

Criterion validity focuses on comparisons of results obtained from SP studies with those obtained from alternative methods that are deemed to elicit true preferences; for example simulated markets in experimental settings or real choice/market setting (Mariel et al. 2021). Criterion validity is strictly related to the vast literature addressing HB (see Table 1).

Hypothetical bias and *ex ante* corrections in DCEs

There is a vast literature exploring the impact that HB has on elicited preferences and WTP in the SP literature. A few meta-analyses showed that hypothetical surveys tend to overestimate WTP with respect to real market settings (e.g., List and Gallet 2001; Murphy et al. 2005; Penn and Hu 2018).

HB in SP studies could be driven by the lack of incentive compatibility of most hypothetical surveys and/or behavioral drivers that may affect participants' responses to the survey question (see discussion proposed by Vossler and Zawojka (2020) on elicitation effects). A number of behavioral drivers has been explored in the literature. First, respondents may be more inclined to state preferences that they think the experimenter wants to hear (i.e., experimenter demand effect) (e.g., Zizzo 2010). Second, respondents may report preferences that they perceive to be more socially acceptable (i.e., social desirability bias) (e.g., Norwood and Lusk 2011). Third, respondents may not perceive their choices as having consequences (i.e., lack of consequentiality), in terms of either the influence of their choices on policy makers' decisions (e.g., Carson and Groves 2007) or on the payment they declared themselves to be willing to pay (e.g., Mitani and Flores 2014). The latter could lead to strategic behavior and free riding. Finally, respondents may also be uncertain about their responses to a discrete choice situation (i.e., preference uncertainty) and make erroneous judgments (e.g., Champ et al. 1997).

This list is far from being exhaustive and we refer interested readers to Loomis (2011, 2014) and Carson, Groves, and List (2014). Back in 1996, Carson et al. reported the lack of a general theory able to explain HB and, despite some attempts to develop such a theory (e.g., Ajzen, Brown, and Carvajal 2004), their claim is still valid up to today.

There are several *ex ante* methods to reduce hypothetical bias, each trying to address one or more of the determinants of HB described above. The implementation of these methods in DCEs is

cross-disciplinary and spans from environmental to health economics, and from energy to agro-food economics (see Table 1).

Cheap talk is a script informing participants about the existence of the HB problem and asking them to respond to the SP survey as if they were in front of a real and binding decision (e.g., Cummings and Taylor 1999). This *ex ante* method generally aims to reduce HB regardless of its determinants. The efficacy of cheap talk is however mixed (e.g., Carlsson, Frykblom, and Lagerkvist 2005; Özdemir, Johnson, and Hauber 2009; Silva et al. 2011; Fifer, Rose, and Greaves 2014; Howard et al. 2017; Penn and Hu 2018; Wuepper, Clemm, and Wree 2019).

Consequentiality is the construction of a survey design that respondents perceive to be consequential in terms of the payment and/or policy implications. The idea is that, if respondents perceive their choices to have an effect on their budget constraints and/or policy makers' decisions, they will make more reliable decisions. This approach was developed for CVM surveys implementing advisory referenda by Carson and Groves (2007) and tested by Carson, Chilton, and Hutchinson (2009) and Vossler and Evans (2009). It was extended to DCEs by Vossler, Doyon, and Rondeau (2012). While this approach provided encouraging results in many fields of applied economics (e.g., Czajkowski et al. 2017; Lewis, Grebitus, and Nayga Jr 2017; Oehlmann and Meyerhoff 2017; Zawojcka, Bartczak, and Czajkowski 2019; Carson et al. 2020), its implementation is most appropriate for public goods.

Honesty priming consists of presenting subjects with a task that indirectly emphasizes the value of honesty among respondents. The honesty priming task is presented to respondents before the DCE. This approach was developed by de-Magistris, Gracia, Nayga Jr. (2013) and there is empirical evidence suggesting that it can mitigate HB in DCE studies (e.g., Bello and Abdulai 2016; Howard et al. 2017). Other studies have asked respondents to read and sign oath scripts in which they swear to tell the truth and provide honest answers during the survey. This approach that was proposed by Jacquemet et al. (2013) empirically reduces HB in SP studies (e.g., Jacquemet et al. 2017). It has been recently implemented in DCE studies (e.g., de-Magistris and Pascucci 2014;

Kemper, Popp, and Nayga Jr. 2017; Mamkhezri et al. 2020). Oath scripts and cheap talk appear to provide the highest reductions of HB when combined (e.g., Jacquemet et al. 2013). Another approach is the use of virtual reality (VR) in DCEs. Fang et al. (2020) found that the use of VR can reduce hypothetical bias particularly for those who do not significantly experience VR discomfort. The use of VR has been proven to also reduce variability in preferences (e.g., Haghani and Sarvi 2019) as well as the asymmetry between WTP and WTA (Bateman et al. 2009). While cheap talk, honesty priming, consequentiality, and oaths have been extensively tested and appear to mitigate, at least to some extent, HB in SP surveys, indirect questioning (IQ) and BTSs have not yet received much attention. We discuss these in more detail in the next subsections, and how they relate to the CMA mechanism.

Table 1

Synthesis of contributions to the DCE literature by discipline^a

Discipline	Reliability	Convergent Validity	Criterion Validity^b
Environmental economics	Bliem, Getzner, and Rodiga-Laßnig 2012 Schaafsma et al. 2014 Czajkowski, Barczak, and Budziński 2016 Rakotonarivo, Schaafsma, and Hockley 2016 Brouwer, Logar, and Sheremet 2017 Matthews, Scarpa, and Marsh 2017	Hanley et al. 1998 Boyle et al. 2001 Caparros, Oviedo, and Campos 2008 Christie and Azevedo 2009	Bateman et al. 2009 Vossler, Doyon, and Rondeau 2012 Czajkowski et al. 2017 Howard et al. 2017 Jacquemet et al. 2017 Carson et al. 2020
Agri-food economics	Mørkbak and Olsen 2015 Rigby, Burton, and Pluske 2016	Asioli et al. 2016 Yangui et al. 2019	Carlsson, Frykblom, and Lagerkvist 2005 De-Magistris, Gracia and Nayga Jr. 2013 de-Magistris and Pascucci 2014 Bello and Abdulai 2016 Lewis, Grebitus, and Nayga Jr. 2017 Kemper, Popp, and Nayga Jr. 2017 Wuepper, Clemm, and Wree 2019 Fang et al. 2020
Health economics	Bryan et al. 2000 Ryan et al. 2006 Skjoldborg, Lauridsen, and Junker 2009 Janssen et al. 2017 Price, Dupont, and Adamowicz 2017	Van der Pol et al. 2008 Ryan and Watson 2009	Özdemir, Johnson, and Hauber 2009

Energy economics	Liebe, Meyerhoff, and Hartje 2012	McNair, Bennett, and Hensher 2011	Oehlmann and Meyerhoff 2017 Zawojkska, Bartczak, and Czajkowski 2019
Transportation economics	Börjesson 2014	Raffaelli et al. 2021	Fifer, Rose, and Greaves 2014 Haghani and Sarvi 2019 Mamkhezri et al. 2020
Cross-sector Studies	Johnstone et al. 2017 Bishop and Boyle 2019 Mariel et al. 2021	Holmes, Adamowicz, and Carlsson 2017 Johnstone et al. 2017 Bishop and Boyle 2019 Lloyd-Smith, Zawojkska, and Adamowicz 2021 Mariel et al. 2021	List and Gallet 2001 Murphy et al. 2005 Johnstone et al. 2017 Penn and Hu 2018 Bishop and Boyle 2019 Mariel et al. 2021

Note: ^a This table is far from providing an exhaustive list of works in the selected areas of research. ^b This column includes also all the literature related to hypothetical bias and ex ante correction methods

Indirect Questioning and Novel Truth Serums

The IQ method goes back to Haire (1950) and involves asking respondents to predict the choice behavior of a third party which is indicated in the indirect question (Fischer and Tellis 1998). In SP studies, IQ (sometimes referred to as inferred valuation (Lusk and Norwood 2009b)), involves asking respondents to predict choices of other respondents or the population of interest, instead of reporting their own private choices. In the context of a DCE, this involves not choosing one's own most preferred amongst a number of options, but estimating the distribution of choices that the population (of interest) would make in each choice situation. This can reduce social desirability bias which is identified as a driver of HB; therefore it is particularly relevant for the elicitation of preferences associated with public goods and/or attributes. In theory, IQ could reduce HB in SP studies (Norwood and Lusk 2011) and there is empirical evidence suggesting that IQ is able to partially reduce HB in DCE studies (e.g., Lusk and Norwood 2009b; Carlsson, Daruvala, and Jaldell 2010; Yadav, van Rensburg, and Kelley 2013; Menapace and Raffaelli 2020; Raffaelli et al. 2021). The implementation of IQ is relatively straightforward and an IQ DCE survey can be conducted in the same manner as a standard DCE survey. Nonetheless, whilst representing an improvement over classic DCEs, IQ is not without problems. Firstly, IQ is still not incentive compatible as respondents have no incentives to provide truthful beliefs about choices at the population level. Further, IQ methods are only able to elicit population level choices (i.e.,

distributions of population choices) and it is not clear whether respondent's beliefs about others' choices correlate with their own individual preferences (Fisher 1993). Thus, whilst IQ may offer some advantages over standard DCE, it is certainly not a panacea.

More recently, new methods to elicit truthful choices were applied to DCEs. The Bayesian Truth Serum (BTS, Prelec 2004) method asks respondents to make their personal choice and also to predict the choice behavior of other participants, just as in the IQ method. The BTS uses a score mechanism that is associated with monetary payoffs and induces respondents to provide truthful personal choices. This is because truthful revelation is a Bayesian Nash Equilibrium (BNE). The BTS' scoring rule consist of two components: i) a "prediction score" that reward respondents' predictions based on their accuracy with respect to the others' choice behavior, and ii) an "information score" that rewards respondents' personal choices based on whether these are surprisingly common, meaning these choices are more frequent than predicted.³ Menapace and Raffaelli (2020) applied the BTS in the context of a DCE by exploring consumers' preferences for more sustainably produced pasta. Respondents were asked to make choices in a standard hypothetical DCE and guess the percentage of respondents choosing each available option in each of the presented choice situations. The BTS's score was associated with a payment rule that should incentivize respondents to make truthful personal choices: the top 30% of respondents in terms of this Bayesian Truth Serum score received a €30 gift voucher.

The BTS has its own limitations. First, the method requires a large sample, where a large sample is defined as a function of the (unknown) prior and is therefore impossible to calculate *a priori* (i.e., it is impossible to know how large a sample is required in advanced). Second, the implementation of the BTS in DCE surveys requires the payment of monetary rewards to respondents which make this approach slightly more difficult to operationalize than a standard DCE survey. Finally, the BTS is more cognitively demanding than a standard DCE survey and this may create fatigue effects that undermine the accuracy of elicited preferences. Further refinements of the BTS either cannot be applied to DCEs, such as the Robust BTS (RBTS) by Witkowski and Parkes

(2012) or have other undesirable properties such as the Divergent BTS (DBTS) by Radanovic and Faltings (2014) which allows for dishonest equilibria where lying is a payoff-dominant strategy (Cvitanic et al. 2019). In this paper, we focus on the ability of the CMa mechanism to overcome some of the limitations of IQ and BTS discussed above, especially with respect to applying the methods in DCE applications.

2. The Choice Matching method applied to the DCE survey

This paper is the first application of the CMa method to a DCE. The CMa method consists of two stages that can be operationalized into two tasks. Task 1 (preference task) is equivalent to a standard hypothetical DCE where respondents are asked to select the most preferred alternative in several choice situations. In task 2 (belief task), each respondent i is asked to predict the choices that all other respondents in their session made in each choice situation k presented in task 1. These predictions are elicited using an incentivized proper scoring rule (i.e., quadratic scoring rule, QSR) and respondents are rewarded depending on how accurate their predictions are at the end of the experiment. This induces respondents to reveal their true beliefs about the choices that other respondents made in task 2.

The key mechanic that makes the DCE presented in task 1 incentive compatible is that at the beginning of the experiment, respondents are informed that:

- i) one choice situation k will be selected at random (i.e., binding choice) at the end of the experiment
- ii) there is a probability $p = [0,1]$ that the prediction that each respondent i made in task 2 regarding the choices that other respondents (j) made in the binding choice situation k in task 1 will be replaced by the average prediction of all other respondents that made the same choice as respondent i in task 1. Therefore, respondents may receive an experimental payoff according to the average predictions of the other respondents who made the same choice as them.

The key assumption that needs to be satisfied to make the DCE in task 1 incentive compatible is referred as impersonal updating. This assumption implies that respondents who made the same choice in choice situation k of task 1 report the same beliefs regarding the choices that other respondents made in choice situation k of task 1. The logic is that respondents prefer that their beliefs expressed in task 2 are replaced by beliefs more similar to their own simply because this maximizes their expected experimental payoff. Therefore, to increase the probability of maximizing their own experimental payoff from task 2, each respondent has an incentive to report truthful choices in task 1. This is true assuming that respondents' beliefs and choices are highly correlated (impersonal updating). In fact, if a respondent makes untruthful choices in task 1, there is a chance that their payoff from stage 2 will be determined by beliefs that differ from their own true belief. In this case, the respondent will not maximize their experimental payoff. Thus, respondents have strict incentives to provide their truthful choices in task 1, despite the preference elicitation not being directly incentivized in task 1.

In theory, the CMA mechanism should fully remove HB and induce demand revelation and hence should represent a marked improvement over the IQ method discussed above. As compared to the Bayesian Truth Serum (BTS), the CMA has the advantages of: i) not being based on any kind of equilibrium concept or requiring cognitively difficult Bayesian updating, ii) using a payoff generating rule which is easier to explain, and iii) being implementable in small groups. Similarly to the BTS, the implementation of the CMA in DCE surveys implies the payment of monetary payoff to respondents. This represents a complication with respect to the design of a standard DCE survey.

3. Methods

Discrete Choice Experiment

In the DCE, respondents were presented with a series of 12 choice situations that each featured two alternative cottage pies (Options A and B) and an 'I prefer neither' opt-out alternative (Option C). Each cottage pie was the same size – 400g – representing a typical

individual portion size (i.e., a ‘ready meal for one’). A cottage pie is made of a layer of mashed potato on top of minced beef in gravy/sauce with some vegetables (onions, carrots, etc.) included. Options A and B were described using four attributes.

The first attribute is related to the presence of antibiotic traces in food. Each cottage pie either had a label indicating that the (cattle derived) ingredients (i.e., beef and dairy) were ‘Raised without Antibiotics’ or else had no such label (implicitly indicating the possible presence of antibiotic residues in the meat and dairy ingredients). This attribute had 2 levels.

The cottage pies were also described according to the saturated fat and salt content using a traffic light system (TLS). The Food Standards Agency (FSA) has implemented a (voluntary) TLS that rates food products as either low, medium or high (represented as green, amber or red, respectively) for their quantities of calories, fat, saturated fat, sugar and salt (FSA 2016). These attributes had 3 levels each. The level of saturated fat is varied as being either low (or green: 1.2g per 100g), medium (or amber: 2.3g per 100g) or high (or red: 6.2g per 100g). These values are within the appropriate range for each TLS level as per the FSA guidelines and are therefore consistent with existing food labelling that consumers are likely to be familiar with. Secondly, the level of salt was varied in a similar manner. Low (or green) salt content corresponds to 0.2g per 100g, medium (or amber) to 1.1g per 100g and finally high (or red) to 2.3g per 100g.

Finally, each cottage pie was associated with one of four prices (attribute levels): £1.50, £2.00, £3.00 or £4.50. This range of prices represents the extent of market prices from, at the low end, a supermarket ‘own brand’ version up to, at the high end, a fancy or gourmet version from an upmarket supermarket.

The 12 choice situations were generated using a D-efficient design which was created using data from a pilot study (ChoiceMetrics 2018)⁴. An example choice situation that was shown to respondents during the instructions can be seen in Figure 1. Respondents were asked to select their most preferred alternative in each choice situation. The order of the choice situations was

randomized across respondents. Since this DCE was hypothetical, respondents do not actually have the chance to receive the cottage pie or any additional payment beyond the £15 participation fee.

Figure 1

An example choice situation shown to respondents in the experimental instructions.

Note: light grey corresponded to yellow in the instructions, mid grey corresponded to green in the instructions, dark grey corresponded to red in the instructions.

Choice-matching discrete choice experiment

As previously mentioned, the CMa mechanism consisted of two tasks. In task 1 (preference task), respondents were exposed to the same procedure used for the classic hypothetical DCE, as described in above. In task 2 (belief task), they were asked to predict the frequency of other participants (in their session) choosing each option (A, B or the opt-out C) in task 1. This was asked for each of the 12 choice situations presented in task 1.

These frequencies were elicited using a quadratic scoring rule (QSR) (Brier 1950 and Murphy and Winkler 1970). Each possible prediction corresponded to a payoff vector of three possible payoffs (see Figure 2). These payoffs were derived using the implementation of QSR for eliciting subjective probability distributions recently developed by Harrison et al. (2017). The QSR rewards respondents for the accuracy of their predictions and penalizes them depending on how frequencies are distributed across the available intervals (Harrison et al. 2017). In each choice situation k , the exact payoff associated with each option j (A, B or C) is calculated as $\pi_{i,j,k} = a + b[(2 * p_{i,j,k}) - \sum_{l=A,B,C} p_{i,l,k}^2]$, where $p_{i,j,k}$ is the frequency assigned by respondent i to a given option j in choice situation k . In our parameterization, $a = b = 5$ giving a minimum payoff of £0 and a maximum of £10. Respondents in the CMa treatment therefore received a payment ranging from £15 to £25.⁵ For example, assume that a respondent i in a session with 10 other respondents

predicted that 5 people chose option A, 4 option B and 1 option C in a given choice situation k (as per figure 2). The payoff obtained by respondent i for each of the options A, B and C is then given by $\pi_{i,A,k} = 5 + 5[(2 * 0.5) - (0.5^2 + 0.4^2 + 0.1^2)] = £7.90$ if option A realizes, to $\pi_{i,B,k} = 5 + 5[(2 * 0.4) - (0.5^2 + 0.4^2 + 0.1^2)] = £6.90$ if option B realizes, and to $\pi_{i,C,k} = 5 + 5[(2 * 0.1) - (0.5^2 + 0.4^2 + 0.1^2)] = £3.90$ if option C realizes.

Figure 2

An example choice situation shown to respondents in the experimental instructions.

Note: light grey corresponded to yellow in the instructions, mid grey corresponded to green in the instructions, dark grey corresponded to red in the instructions.

Respondents were provided with the following pieces of information before taking part in the experiment. This information was related to the procedures used to calculate their additional earnings.

Information 1. Respondents were told they will take part in two tasks in the following order: the preference task (task 1) and the belief task (task 2).

Information 2. They were told that one choice situation was to be drawn at random to be payoff relevant and this was referred to as the binding choice situation. This was illustrated to respondents as a numbered ball (from 1 to 12) drawn from a bucket.

Information 3. They were also informed that earnings depended on: i) the reported frequency of respondents choosing each option l (A, B or C) in the binding choice situation k in task 1, and ii) the observed frequency of respondents choosing each option l (A, B or C) in the binding choice situation k in task 1. In particular, respondents' payoff depended on a random draw from a bucket containing labelled balls. These balls were labelled as A, B, and C. The proportion of A-, B- and C-labelled balls in the bucket was equivalent to the observed frequency of respondents

choosing option A, B or C in the binding choice situation k in task 1. The final earnings for each respondent were equal to the payoff calculated using the QSR for the randomly drawn letter. Consider the example reported in Figure 2. Respondent i would earn £7.90 if an A-labelled ball was randomly drawn from the bucket, while they would earn £6.90 if a B-labelled ball was picked from the bucket and finally £3.90 if a C-labelled ball was randomly drawn.

Information 4. Respondents were also told that there was a chance (70% in our experiment ⁶) that their payoffs were calculated based on the average predictions of the other respondents who preferred the same option as themselves in the binding situation in task 1, rather than calculated based on their predictions. This might affect their payoff. As an example, suppose respondent i preferred Option A in the binding choice situation in task 1. There is a 70% chance that respondent i 's predictions, reported in task 2, regarding the frequency of respondents choosing option A, B, or C in the binding choice situation will be replaced by the average predictions of all the other respondents who also preferred Option A in task 1.⁷ Whether respondent i 's prediction was to be replaced was decided by an independent random draw (illustrated to respondents as the roll of a 10-sided dice). If the outcome of the roll was between 1 and 3, respondent i 's prediction was used to calculate the payoff from task 2 (QSR). If the outcome of the roll was between 4 and 10, the average predictions of the other respondents who preferred the same option as respondent i in the binding situation in task 1 was used to calculate respondent i 's payoff.

Sample and experimental design

Our sample consists of 130 consumers living in or around Belfast (Northern Ireland, United Kingdom).⁸ The study was advertised in a number of locations (including digital channels) and described simply as a food choice study. Ages ranged from 19 to 74 (the average age was 33) and 65% of respondents were female.

As previously discussed, our sample was randomly split between our two treatment conditions, 66 respondents took part in the treatment featuring the DCE supplemented by the

CMA mechanism (hereafter, ‘CMA treatment’) and the remaining 64 respondents took part in the DCE baseline control treatment (‘DCE treatment’). Sessions in the CMA treatment ranged in size from 10 to 14 respondents. Sessions in the DCE treatment ranged in size from 6 to 14 respondents. Participants were randomly assigned to sessions. All sessions took place at the Institute for Global Food Security at Queen’s University Belfast in April 2019. All sessions were programmed and run using z-Tree (Fischbacher 2007). The study was granted full ethical approval by the Faculty of Medicine, Health and Life Sciences Ethical Review Board at Queen’s University Belfast.

In both treatments, respondents began by answering two questions relating to how hungry or full they felt at that moment, rated on a 7-point Likert scale. Subsequently, using a D-efficient design in both treatments, respondents made their choices in the 12 choice situations presented in a standard DCE survey. In the CMA treatment only, respondents then expressed their beliefs regarding other respondents’ choices in the standard DCE incentivized via the QSR. To familiarize with the CMA procedures, respondents were exposed to a practice involving both the preference and the belief tasks. Finally, all respondents in both treatments completed a questionnaire that asked about demographics, behavior, and preferences related to cottage pies (including expected taste for every combination of saturated fat and salt content), wider shopping habits and knowledge of antibiotics and anti-microbial resistance.

All participants received a £15 show-up fee prior to making any of their choices (or hearing any of the instructions) in the experiment. An overview of each of the treatments can be seen in Table 2.⁹

Table 2

Layout of steps in each experimental treatment

DCE Treatment		CMA Treatment	
Step	Description	Step	Description
1	Pre-questionnaire: Respondents answer questions about how hungry and how full they feel	1	Pre-questionnaire: Respondents answer questions about how hungry and how full they feel
-	N/A	2	Practice: Respondents take part in a practice Preference Task and Belief Task
2	Preference Task: Respondents make their choices in the 12 choice situations of the DCE (Direct Questioning (DQ), individual level choices elicited)	3	Preference Task: Respondents make their choices in the 12 choice situations of the DCE (DQ, individual level choices elicited)
-	N/A	4	Belief Task: Respondent make their predictions about other respondents' choice behavior (IQ, population level choices elicited)
3	Post-questionnaire: Final questionnaire	5	Post-questionnaire: Final questionnaire
-	N/A	6	Payoff disclosure: Respondents are informed regarding their additional payoff
-	N/A	7	Payment Respondents are paid the additional payment (if any)

4. Testing the Choice Matching's ability to reduce hypothetical bias

Econometric models and testable hypotheses

In this paper, we test the validity of CMA using a criterion validity paradigm. Specifically, we compare marginal WTP (mWTP) for each attribute characterizing the cottage pie across treatments. We expect that mWTPs elicited via CMA will be lower than those elicited via the hypothetical DCE which can potentially suffer from HB. Also, reliability of mWTPs elicited via CMA is compared with that of mWTPs elicited via the hypothetical DCE. Specifically, we compared:

- i) Standard deviations of estimated mWTPs across treatments as a measure of reliability as suggested by Kealy (1990). We expect that standard deviations associated with CMA will be lower than those elicited via the DCE. We acknowledge that using standard deviation as an indicator of reliability of estimated mWTPs is not the norm in the choice modelling literature. These are usually used as an indicator of preference heterogeneity (Hensher and Greene 2003).
- ii) Standard errors of estimated coefficients in our choice models following Bishop and Boyle (2019). As larger standard errors signal lower reliability (and vice versa), we expect that standard errors will be lower in the CMA treatment than in the hypothetical DCE.
- iii) Scale parameter of estimated choice models as suggested by Liebe, Meyerhoff, and Hartje (2012). We expect that the scale parameter (error variance) will be higher (lower) in the CMA treatment than in the hypothetical DCE.

To this end, we used a two-step procedure. First, we estimate the mean and standard deviation of *mWTPs* by estimating two RPL models (or mixed logit) in WTP space (Train and Weeks 2005): Model 1 is estimated using data from the CMA treatment, Model 2 using data from

the DCE treatment. Second, we compare the mean and standard deviations of $mWTP$ s across treatments using Poe, Giraud, and Loomis' convolution approach (2005).¹⁰

In the first step, Random Utility Models (RUMs) were used to model choice data (McFadden 1973). RUMs assume that the utility that participant i attaches to each alternative j in each choice situation k is split into two parts; $V_{i,j,k}$, the part of the utility observed by the researcher, and $\varepsilon_{i,j,k}$, which cannot be observed by the researcher, so that, $U_{i,j,k} = V_{i,j,k} + \varepsilon_{i,j,k}$. RPL models were estimated in WTP space because this estimation procedure provides a number of advantages with respect to standard estimation in preference space. First, it allows direct estimation of $mWTP$ for non-price attributes. In WTP space models, the utility is re-arranged such that estimated coefficients related to non-price attributes represent $mWTP$ for such attributes. Second, estimation in WTP space mitigates the confounding of variation in scale (i.e. the standard deviation of the unobserved part of the utility) and WTP (Train and Weeks 2005) which is instead an issue in models estimated in preference space. Third, many studies have shown that models in WTP space fit data better than those in preference space (e.g., Thiene and Scarpa 2008; Hole and Kolstad 2012). This estimation approach was recently adopted in studies investigating consumers' preferences for food products (e.g., Lin, Ortega, and Caputo 2019; Macdiarmid et al. 2021).

The general specification of the indirect utility function of the RPL models estimated in WTP space is specified as:

$$V_{i,j,k} = -\lambda_i PRICE_{i,j,k} + (\lambda_i \omega_i) \mathbf{x}_{i,j,k} \quad [1]$$

In Equation 1, $\lambda_i = \alpha_i / \mu_i$, where α_i indicates participants' preferences for the price of the cottage pie $PRICE_{i,j,k}$ and μ_i is the scale parameter (the standard deviation of the unobserved part of the utility). The coefficient vector $\omega_i = \beta_i / \alpha_i$ is the ratio of the vector of coefficients β_i that are associated to the vector of non-price attributes $\mathbf{x}_{i,j,k}$ and the coefficient α_i . The vector of coefficients β_i indicates preferences for the vector of non-price attributes $\mathbf{x}_{i,j,k}$, while the vector of coefficients ω_i

indicates the vector of $mWTPs$ associated with the vector of non-price attributes $\mathbf{x}_{i,j,k}$.

The vector ω_i is composed of the following coefficients. The coefficients $\omega_{FAT_A,i}$ and $\omega_{FAT_G,i}$ indicate subjects' $mWTP$ for pies that are amber and green in saturated fat (FAT_A and FAT_G , respectively) compared to pies that are red in saturated fat (FAT_R). The coefficients $\omega_{SALT_A,i}$ and $\omega_{SALT_G,i}$ indicates subjects' $mWTP$ for pies that are amber and green in salt ($SALT_A$ and $SALT_G$, respectively) compared to pies that are red in salt ($SALT_R$). The coefficient $\omega_{ANT,i}$ refers to pies that are made of beef and dairy products produced from animals that were 'Raised without Antibiotics'. The coefficient $\omega_{OPT-OUT,i}$ indicates subjects' preferences for the opt-out alternative with respect to the cottage pie alternatives. The coefficients $\omega_{FAT_A,i}$, $\omega_{FAT_G,i}$, $\omega_{SALT_A,i}$, $\omega_{SALT_G,i}$, $\omega_{ANT,i}$ and $\omega_{OPT-OUT,i}$ and are all assumed to be normally distributed with means and standard deviations to be estimated. The coefficient α_i indicates subjects' preferences for the price of pies (PR) and is modelled as a random parameter following a log-normal distribution with mean and standard deviation to be estimated.

In the second step, we used Poe et al.'s convolution approach (2005) to test differences in the distribution of estimated coefficients between Model 1 (CMA) and Model 2 (DCE). Specifically, we used parametric bootstrapping techniques (i.e., Krinsky and Robb 1986) to generate 1,000 bootstrapped values for each pair of coefficient distributions and calculated 1,000,000 differences between the two bootstrapped distributions. The full set of hypotheses that are tested is reported in Table 3.

Table 3
Description and interpretation of testable hypotheses

Validity		Reliability	
Null Hypothesis (H ₀)	Interpretation	Null Hypothesis	Interpretation
$\omega_{OPT_OUT, MEAN, DCE} \geq \omega_{OPT_OUT, MEAN, CMa}$	Rejecting H ₀ means that mWTP for the opt-out alternative is higher in the CMa than in the DCE. The CMa reduces HB as compared to the DCE	$\omega_{OPT_OUT, SD, DCE} \leq \omega_{OPT_OUT, SD, CMa}$	Rejecting H ₀ means that standard deviations of price and non-price coefficients are lower in the CMa than in the DCE. The CMa provides less dispersed coefficient values than the DCE
$\omega_{FAT_A, MEAN, DCE} \leq \omega_{FAT_A, MEAN, CMa}$ $\omega_{FAT_G, MEAN, DCE} \leq \omega_{FAT_G, MEAN, CMa}$ $\omega_{SALT_A, MEAN, DCE} \leq \omega_{SALT_A, MEAN, CMa}$ $\omega_{SALT_G, MEAN, DCE} \leq \omega_{SALT_G, MEAN, CMa}$	Rejecting H ₀ means that mWTP for the non-price attributes are greater in the DCE than in the CMa. The CMa reduces HB as compared to the DCE	$\omega_{FAT_A, SD, DCE} \leq \omega_{FAT_A, SD, CMa}$ $\omega_{FAT_G, SD, DCE} \leq \omega_{FAT_G, SD, CMa}$ $\omega_{SALT_A, SD, DCE} \leq \omega_{SALT_A, SD, CMa}$ $\omega_{SALT_G, SD, DCE} \leq \omega_{SALT_G, SD, CMa}$	
$\omega_{PRICE, MEAN, DCE} \geq \omega_{PRICE, MEAN, CMa}$	Rejecting H ₀ means that the impact of price on respondents' decisions is lower in the DCE than in the CMa. The CMa reduces HB as compared to the DCE	$\omega_{PRICE, SD, DCE} \leq \omega_{PRICE, SD, CMa}$	
		$\tau_{DCE} \geq \tau_{CMa}$	Rejecting H ₀ means that error variance in the CMa is lower than in the DCE. The CMa provides more deterministic choices than the DCE

Results

Results from the estimation of Models 1 and 2 are reported in Table 4. Summary statistics of the choices made in the two treatment groups are provided in the online supplementary appendix B. Potential differences among the two sub-samples were investigated using a logit sample selection model, non-parametric Kolmogorov-Smirnov tests, and parametric t-tests. We did not find any substantial difference in a set of key variables (e.g., gender, age, hunger level, income, taste expectation for the pies, etc.).¹¹

Our results in Table 5 suggest that the distributional means of our coefficients are not statistically different across groups, meaning that there is no evidence that *mWTPs* elicited via CMA are lower than those elicited via the DCE. After all, it may be that *mWTPs* elicited via hypothetical DCE already have an acceptable level of validity. It is common knowledge that SP studies valuing private goods are less affected by hypothetical bias than SP studies valuing public goods (List and Gallet 2001; Murphy et al. 2005; McFadden and Train 2017). In a more recent meta-analysis, Penn and Hu (2018) found that DCE and referendum formats generate significantly lower HB than open-ended, payment card and dichotomous choice CVM studies. Therefore, we conclude that the *mWTP* estimated from the CMA method are as valid as those estimated from the DCE.

Concerning reliability, we found that CMA provides less dispersed distribution for four (out of seven) attributes in our empirical application (i.e., *OPT-OUT*, *FAT_A*, *FAT_G*, *SALT_G* and *ANT*), which suggests that *mWTPs* elicited via the CMA may be more reliable than those elicited using the hypothetical DCE survey.¹² This finding is confirmed by the fact that standard errors of estimated coefficients are consistently smaller in the CMA than in the DCE treatment (see Table 4). However, we found that the scale parameter τ and error variance are not statistically significantly different in the two groups (Table 5). Two indicators of reliability (out of three) suggest that CMA provides more reliable results than the hypothetical DCE survey, indicating that the CMA has the potential to improve the reliability of estimated welfare measures.

Table 4Random Parameter Logit models estimated in WTP space^a

	Model 1 (CMa)	Model 2 (DCE)
Dep.Var.:	<i>CHOICE</i>	<i>CHOICE</i>
	Coefficient	Coefficient
$\omega_{OPT_OUT, MEAN}$	1.342*** (0.391)	1.865*** (0.477)
$\omega_{FAT_A, MEAN}$	1.320*** (0.198)	1.569*** (0.353)
$\omega_{FAT_G, MEAN}$	2.406*** (0.311)	2.584*** (0.464)
$\omega_{SALT_A, MEAN}$	0.783*** (0.188)	0.459* (0.251)
$\omega_{SALT_G, MEAN}$	1.571*** (0.188)	1.633*** (0.316)
$\omega_{ANT, MEAN}$	0.861*** (0.141)	0.879*** (0.179)
α_{MEAN}	-0.981*** (0.135)	-1.344** (0.566)
$\omega_{OPT_OUT, SD}$	2.297*** (0.384)	4.698*** (0.862)
$\omega_{FAT_A, SD}$	0.352* (0.194)	0.653*** (0.224)
$\omega_{FAT_G, SD}$	0.673** (0.257)	1.970*** (0.389)
$\omega_{SALT_A, SD}$	0.881*** (0.223)	1.602*** (0.328)
$\omega_{SALT_G, SD}$	0.490 (0.345)	2.345*** (0.450)
$\omega_{ANT, SD}$	0.919*** (0.206)	2.579*** (0.504)
α_{SD}	0.800** (0.398)	1.741 (1.326)
τ	0.611*** (0.160)	0.993*** (0.232)
Subjects	66	64
Observations	2,376	2,304
Log Likelihood	-648.447	-645.453
BIC	1413.492	1407.042

Note: ^a Robust standard errors in brackets

*p<0.10; **p<0.05; ***p<0.01

Table 5.

Comparisons of distributional means and standard deviation of estimated coefficients across treatments using the Poe et al.'s test (2005)^a

Coefficients	CMa	DCE	H₀ (Null Hypothesis)	P-value
$\omega_{OPT_OUT,MEAN}$	1.345 (0.677; 1.975)	1.875 (1.037; 2.648)	$\omega_{OPT_OUT,MEAN,DCE} \geq \omega_{OPT_OUT,MEAN,CMa}$	0.802
$\omega_{FAT_A,MEAN}$	1.317 (1.004; 1.636)	1.584 (1.002; 2.157)	$\omega_{FAT_A,MEAN,DCE} \leq \omega_{FAT_A,MEAN,CMa}$	0.252
$\omega_{FAT_G,MEAN}$	2.408 (1.897; 2.915)	2.585 (1.795; 3.343)	$\omega_{FAT_G,MEAN,DCE} \leq \omega_{FAT_G,MEAN,CMa}$	0.376
$\omega_{SALT_A,MEAN}$	0.788 (0.479; 1.107)	0.465 (0.048; 0.874)	$\omega_{SALT_A,MEAN,DCE} \leq \omega_{SALT_A,MEAN,CMa}$	0.845
$\omega_{SALT_G,MEAN}$	1.573 (1.253; 1.880)	1.645 (1.089; 2.170)	$\omega_{SALT_G,MEAN,DCE} \leq \omega_{SALT_G,MEAN,CMa}$	0.419
$\omega_{ANT,MEAN}$	0.866 (0.636; 1.093)	0.879 (0.599; 1.185)	$\omega_{ANT,MEAN,DCE} \leq \omega_{ANT,MEAN,CMa}$	0.479
$\omega_{PRICE,MEAN}$	-1.033 (-1.540; -0.674)	-1.556 (-2.988; -0.704)	$\omega_{PRICE,MEAN,DCE} \leq \omega_{PRICE,MEAN,CMa}$	0.754
$\omega_{OPT_OUT,SD}$	2.300 (1.674; 2.934)	4.722 (3.184; 6.119)	$\omega_{OPT_OUT,SD,DCE} \leq \omega_{OPT_OUT,SD,CMa}$	0.007
$\omega_{FAT_A,SD}$	0.358 (0.031; 0.686)	0.662 (0.289; 1.033)	$\omega_{FAT_A,SD,DCE} \leq \omega_{FAT_A,SD,CMa}$	0.156
$\omega_{FAT_G,SD}$	0.671 (0.221; 1.144)	1.975 (1.305; 2.625)	$\omega_{FAT_G,SD,DCE} \leq \omega_{FAT_G,SD,CMa}$	0.005
$\omega_{SALT_A,SD}$	0.886 (0.506; 1.240)	1.613 (1.050; 2.154)	$\omega_{SALT_A,SD,DCE} \leq \omega_{SALT_A,SD,CMa}$	0.363
$\omega_{SALT_G,SD}$	0.469 (0.106; 1.045)	2.355 (1.603; 3.121)	$\omega_{SALT_G,SD,DCE} \leq \omega_{SALT_G,SD,CMa}$	0.001
$\omega_{ANT,SD}$	0.922 (0.554; 1.727)	2.603 (1.691; 3.416)	$\omega_{ANT,SD,DCE} \leq \omega_{ANT,SD,CMa}$	0.002
$\omega_{PRICE,SD}$	0.847 (0.470; 1.439)	1.840 (0.533; 4.756)	$\omega_{PRICE,SD,DCE} \leq \omega_{PRICE,SD,CMa}$	0.203
τ	0.721 (0.462; 0.974)	1.010 (0.627; 1.417)	$\tau_{DCE} \leq \tau_{CMa}$	0.845

Note: ^a The 5% and the 95% percentiles are in brackets

5. Testing impersonal updating and result robustness

The key assumption behind the CMa method is impersonal updating: respondents with similar choices should have similar beliefs about other respondents' choices. If this assumption does not hold, the incentive compatibility of the CMa method may be weakened.

We test whether impersonal updating is satisfied empirically using Pearson's χ^2 tests. Specifically, in each session t , for each choice situation k , we tested whether subjects who chose the same alternative j (A, B or C) in the preference task (task 1) had equivalent beliefs regarding the number of subjects choosing option A, B and C in the preference task (task 1). Beliefs were reported by respondents in the belief task (task 2). This is our null hypothesis (H_0) to test. Rejection of this null hypothesis suggests that impersonal updating was not satisfied. We conducted 18 tests per choice situation k , 3 for each session t (we had 6 sessions). Overall, we conducted 216 tests as we had 12 choice situations. Results from these tests are reported in the online supplementary appendix E. We found that approximately 54% of tests did not reject the null hypothesis of impersonal updating, suggesting that impersonal updating is satisfied in just over half of the cases.¹³ We discuss the possible ramifications of this result in the next section.¹⁴

6. Conclusions

DCEs are arguably the most popular stated preference method used by applied economists and finding ways to elicit reliable preferences is obviously very important. This paper represents the first empirical application of the choice matching (CMa) method as well as the first empirical tests of its validity and reliability. The CMa method was recently developed by Cvitanic et al. (2019) to elicit honest responses using any type of discrete choice question and represents a refinement of Prelec's Bayesian Truth Serum (2004).

We conducted a lab experiment involving two experimental treatments. Part of the sample was exposed to a standard hypothetical DCE survey, the other part to the CMa method applied to a hypothetical DCE survey. The DCE was designed to elicit consumers' preferences for antibiotic residue presence as well as salt and saturated fat content in cottage pies, a popular British dish which is available as a ready meal on many supermarkets' shelves.

Random parameter logit models (mixed logit) were estimated in WTP space. Our results show that the means of estimated mWTPs for cottage pie attributes do not differ across treatments, indicating that the CMA is as valid as the hypothetical DCE survey. Standard deviations of the estimated mWTPs and the standard errors of the estimated coefficients are significantly lower in the CMA treatment than in the DCE treatment. In contrast, the error variance is not statistically significantly different in the two groups. While these results indicate that the DCE with CMA can be more reliable than the conventional DCE, further research is needed on the topic considering that only two reliability criteria (out of three) support this argument and the standard deviations of the mWTPs may simply signal unobserved preference heterogeneity. We acknowledge that results on validity may be driven by the private nature of the good under investigation. Previous studies have shown that HB is more of a problem when the good under valuation is public (e.g., McFadden and Train 2017; Penn and Hu 2018). Future research could explore the performances of the CMA approach in terms of validity and reliability when the nature of the good under valuation is public or quasi-public. This investigation would be particularly beneficial for disciplines such as environmental and health economics that often focus on valuing these types of goods.

A test of the impersonal updating assumption was conducted to explore whether this crucial assumption for the functioning of the CMA mechanism was satisfied or not. We found that this assumption is only partially satisfied in our sample. This may cast some doubts on the applicability of the CMA method to DCEs and further research is needed to test whether the benefits of using the CMA in stated preference research outweighs the higher cognitive burden that respondents are exposed to in the CMA mechanism. It is possible for example that fatigue effects may undermine the empirical applicability of the CMA for stated preference research. These trade-offs could also be explored by comparing the performance of CMA in controlled lab experiments with more standard subject pools (i.e., students) and field experiments involving the general public.

The potential implementation of CMA in studies that are not conducted in the lab, including those with larger and more representative samples of the population of interest, could be another

important aspect to consider in future research. High levels of internet access (even in remote and rural areas), the availability of many software and platforms to conduct online experiments and the existence of many companies offering consumer panels and sampling services at large scale have increased the potential to conduct economic experiments online. The Covid-19 pandemic has also given more impetus to the use of such online experiments. Recent studies have shown that data quality from online economic experiments is adequate and reliable (e.g., Arechar, Gächter, and Molleman 2017). These considerations allow us to be optimistic about the use of the CMa approach applied to choice-based SP methods outside of the lab with larger and more representative samples. Future research should also explore the validity and reliability of the CMa method in such settings to test the robustness of our findings.

Finally, a limitation of the study is our test of validity, we do not have any real market data to compare our results with. Unfortunately, there were no cottage pies with exactly the same range of attributes as those used in the survey currently available on the UK market when the study was conducted. However, future research could easily perform such a test of validity by using the CMa method to value a good that is currently available in the market. Nevertheless, our empirical application shows that the CMa method can provide more reliable estimates than hypothetical DCE surveys; hence we conclude that CMa could be a promising method that should be further tested not just in DCEs but also in other stated preference elicitation formats, such as the dichotomous choice contingent valuation method, payment card formats, and multiple price list formats. Our hope is that this paper will encourage other researchers to further test the merits of the CMa method in stated preference studies.

Acknowledgement

We thank Chloe McCallum for her help in organizing sessions and running the experiment at Queen's University Belfast. We thank two anonymous reviewers and the editor for comments on an earlier draft of this paper.

Endnotes

¹ In this paper, we use the Harrison and List's (2004) categorization of field experiments.

² Internal validity is the ability to draw robust causal conclusions (Loewenstein 1999).

³ The formula for the Bayesian Truth Serum score is given by $u^r = \sum_{k=1}^m x_k^r \log \frac{\bar{x}_k}{y_k} + \sum_{k=1}^m \bar{x}_k \log \frac{y_k}{x_k}$

where respondent r chooses among $k = 1, \dots, m$ options. x_k^r is then a dummy indicating whether respondent r chose option k ($x_k^r = 1$) or not ($x_k^r = 0$). \bar{x}_k is the proportion of respondents choosing option k . y_k^r is the predicted frequency of respondents choosing option k by respondent r . \bar{y}_k is the average predicted frequency of choosing option k .

⁴ The pilot was conducted with 25 respondents in March 2019. These were randomly recruited among academic and non-academic members of staff at Queen's University Belfast.

⁵ Both treatments took less than 1 hour to complete on average. The minimum wage per hour in the UK in 2019 was £8.21 (UK Government 2019). Therefore, compensation of between £15 and £25 represents a reasonable monetary incentive to motivate participants in their decision making.

⁶ The decision to fix this probability to 70% is due to the fact that we wanted to avoid using a 50% chance since participants may perceive everything as being random (i.e. 'a coin flip') and we didn't want the chance of beliefs being replaced to be so high that participants viewed their own beliefs as being inconsequential in practice. We acknowledge this parameter may influence results. Future research could investigate whether varying this chance influences the properties of the choice matching approach.

⁷ In the case that a participant was the only one to choose a given option – and therefore that there are no existing beliefs to replace the participants with, then the participant automatically gets an additional payment of £0, as per Cvitanić et al. (2019).

⁸ Summary statistics of the S-error were calculated using the software Ngene (ChoiceMetrics 2018): mean = 18.325, standard deviation = 4.022, median=17.720, minimum=11.104 and maximum = 40.949. These suggested that our sample is large enough for the design used in the study.

⁹ The full set of experimental instructions are available in the online supplementary appendix A.

¹⁰ We used parametric bootstrapping techniques (i.e. Krinsky and Robb 1986) to generate 1,000 bootstrapped values for each estimated coefficient and calculate 1,000,000 differences between the two bootstrapped distributions.

¹¹ Results from these analyses are presented in the online supplementary appendix C. Only age was statistically different at the 5% significance level between the two groups.

¹² This is evident in Figure D.1 in the online supplementary appendix D. To test robustness of results, we also estimated models 1 and 2 in preference space and used the Poe et al.'s (2005) convolution approach to test differences in coefficients across treatment groups (CMA and DCE). Results from these analyses show the robustness of our results and are provided in the online supplementary appendix D.

¹³ A limitation of the strategy used to test the impersonal updating assumption is that it relies on the fact that the CMA elicit truthful choices. A proper test of whether elicited choices are truthful (and therefore impersonal updating is satisfied) is only possible if revealed choices are available for the goods under investigation. We thank an anonymous Referee for highlighting the issue.

¹⁴ As pointed out by an anonymous Referee, the key to whether impersonal updating is sufficient for incentive compatibility is the “closeness” of the beliefs. Broadly, it would be sufficient for the beliefs of those who made similar choices to be closer than the beliefs of those who made different choices. Hence, an alternative approach to explore whether the impersonal updating assumption is satisfied would be testing whether beliefs of respondents who made similar choices are more homogenous than beliefs of respondents who made different choices. This would represent a less strict approach to test the impersonal updating assumption.

References

- Adamowicz, Wiktor L. 2004. "What's it worth? An examination of historical trends and future directions in environmental valuation." *Australian Journal of Agricultural and Resource Economics* 48: 419-443.
- Ajzen, Icek, Thomas C. Brown, and Franklin Carvajal. 2004. "Explaining the Discrepancy Between Intentions and Actions: The Case of Hypothetical Bias in Contingent Valuation." *Personality and Social Psychology Bulletin* 30 (9): 1108-1121.
- Arechar, Antonio A., Simon Gächter, and Lucas Molleman. 2018. "Conducting interactive experiments online." *Experimental Economics* 21: 99-131.
- Asioli, Daniele, T. Næs, A. Øvrum, and V.L. Almli. 2016. "Comparison of rating-based and choice-based conjoint analysis models. A case study based on preferences for iced coffee in Norway." *Food Quality and Preference* 48 (A): 174-184.
- Bateman, Ian J., Brett H. Day, Andrew P. Jones, and Simon Jude. 2009. "Reducing gain-loss asymmetry: A virtual reality choice experiment valuing land use change." *Journal of Environmental Economics and Management* 58 (1): 106-118.
- Bello, Muhammad, and Awudu Abdulai. 2016. "Impact of Ex-Ante Hypothetical Bias Mitigation Methods on Attribute Non-Attendance in Choice Experiments." *American Journal of Agricultural Economics* 98 (5): 1486-1506.
- Bishop, Richard C., and Kevin J. Byole. 2019. "Reliability and Validity in Nonmarket Valuation." *Environmental and Resource Economics* 72: 559-582.
- Bliem, Markus, Michael Getzner, and Pietra Rodiga-Laßnig. 2012. "Temporal stability of individual preferences for river restoration in Austria using a choice experiment." *Journal of Environmental Management* 103: 65-73.
- Boyle, Kevin, J., Thomas P. Holmes, Mario F. Teisl, and Brian Roe. 2001. "A comparison of conjoint analysis response formats." *American Journal of Agricultural Economics* 83: 441-454.

- Börjesson, Maria 2014. "Inter-temporal variation in the travel time and travel cost parameters of transport models." *Transportation* 41: 377–396.
- Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1-3.
- Bryan, Stirling, Lisa Gold, Rob Sheldon, and Martin Buxton. 2000. "Preference measurement using conjoint methods: an empirical investigation of reliability." *Health Economics* 9(5): 385–395.
- Brouwer, Roy, Ivan Logar, and Oleg Sheremet. 2017. "Choice consistency and preference stability in test-retests of discrete choice experiment and open-ended willingness to pay elicitation formats." *Environmental Resource Economics* 68: 729–751.
- Camerer, Colin 1995. "Individual decision making." In *The handbook of experimental economics*, ed. John H., Kagel, and Alvin E. Roth, 357-703. Princeton: Princeton University Press.
- Camerer Colin, 1999. "Behavioral economics: Reunifying psychology and economics." *Proceedings of National Academy of Sciences* 96: 10575–10577.
- Campbell, Danny, Marco Boeri, Edel Doherty, and W. George Hutchinson. 2015. "Learning, fatigue and preference formation in discrete choice experiments." *Journal of Economic Behavior & Organization* 119: 345-363.
- Caparros, Alejandro, José L. Oviedo, and Pablo Campos. 2008. "Would you choose your preferred option? Comparing choice and recoded ranking experiments." *American Journal of Agricultural Economics* 90 (3): 843-855.
- Carlsson, Fredrik, Dinky Daruvala, and Henrik Jaldell. 2010. "Do You Do What You Say or Do You Do What You Say Others Do?" *Journal of Choice Modelling* 3 (2): 113-133.
- Carlsson, Fredrik, Peter Frykblom, and Carl Johan Lagerkvist. 2005. "Using Cheap Talk as a Test of Validity in Choice Experiments." *Economics Letters* 89: 147-152.

- Carson, Katherine Silz, Susan M. Chilton, and W. George Hutchinson. 2009. "Necessary conditions for demand revelation in double referenda." *Journal of Environmental Economics and Management* 57 (2): 219-225.
- Carson, Katherine Silz, Susan M. Chilton, W. George Hutchinson, and Riccardo Scarpa. 2020. "Public resource allocation, strategic behavior, and status quo bias in choice experiments." *Public Choice* 185: 1–19.
- Carson, Richard T., Nicholas E. Flores, Kerry M. Martin, and Jennifer L. Wright. 1996. "Contingent valuation and revealed preferences methodologies: Comparing the estimates for quasi-public goods." *Land Economics* 72 (1): 80– 99.
- Carson, Richard T., and Theodore Groves. 2007. "Incentive and informational properties of preference questions." *Environmental and Resource Economics* 37: 181-210.
- Carson, Richard T., Theodore Groves, T., and John A. List. 2014. "Consequentiality: A theoretical and experimental exploration of a single binary choice." *Journal of the Association of Environmental and Resource Economics* 1: 171-207.
- Champ, Patricia A., Richard C. Bishop, Thomas C. Brown, and Daniel W. McCollum. 1997. "Using donation mechanisms to value non use benefits from public goods." *Journal of Environmental Economics and Management* 3: 151–162.
- Christie, Mike, and Christopher D. Azevedo. 2009. "Testing the consistency between standard contingent valuation, repeated contingent valuation and choice experiments." *Journal of Agricultural Economics* 60 (1): 154-170.
- Cerroni, Simone, Verity Watson, Dimitrios Kalentakis, and Jennie I. Macdiarmid. 2019. "Value-elicitation and value-formation properties of discrete choice experiments and experimental auctions." *European Review of Agricultural Economics* 46 (1): 3-27.
- ChoiceMetrics. 2018. Ngene 1.1 *User manual and reference guide*. Available at <http://www.choicemetrics.com/download.html> (accessed May 23, 2018).

- Cummings, Ronald G., and Laura O. Taylor. 1999. "Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method." *American Economic Review* 89: 649-665.
- Cvitanić, Jakša, Dražen Prelec, Blake Riley, and Benjamin Tereick. 2019. "Honesty via Choice-Matching." *American Economic Review: Insights* 1 (2): 179-192.
- Czajkowski, Mikołaj, Anna Barczak, and Wiktor Budziński. 2016. "Preference and WTP stability for public forest management." *Forest Policy and Economics* 71:11–22.
- Czajkowski, Mikołaj, Christian A. Vossler, Wiktor Budziński, Aleksandra Wiśniewska, and Ewa Zawojka, 2017. "Addressing empirical challenges related to the incentive compatibility of stated preferences methods." *Journal of Economic Behavior & Organization* 142: 47-63.
- Day, Brett, Ian J. Bateman, Richard T. Carson, Diane Dupont, Jordan J. Louviere, Sanae Morimoto, Riccardo Scarpa, and Paul Wang. 2012. "Ordering effects and choice set awareness in repeat-response stated preference studies." *Journal of Environmental Economics and Management* 63 (1): 73-91.
- de-Magistris, Tiziana, Azucena Gracia, and Rofolfo M. Nayga Jr. 2013. "On the use of honesty priming tasks to mitigate hypothetical bias in choice experiments." *American Journal of Agricultural Economics* 95 (5): 1136-1154.
- de-Magistris, Tiziana, and Stefanpo Pascucci. 2014. The effect of the solemn oath script in hypothetical choice experiment survey: A pilot study. *Economics Letters* 123 (2): 252-255.
- Fang, Di, Rodolfo M. Nayga Jr., Grant West, Claudia Bazzani, Wei Yang, Benjamin C. Lok, Charles Levy, and Heather A. Snell, 2021. "On the Use of Virtual Reality in Mitigating Hypothetical Bias in Choice Experiments." *American Journal of Agricultural Economics*, 103 (1): 142-161.
- Fifer, Simone, John Rose, and Stephen Greaves. 2014. "Hypothetical bias in Stated Choice Experiments: Is it a problem? And if so, how do we deal with it?" *Transportation Research Part A: Policy and Practice* 61: 164-177.

Fischbacher, Urs. 2007. “z-Tree: Zurich toolbox for ready-made economic experiments.”

Experimental Economics 10 (2): 171-178.

Fisher, Robert J. 1993. “Social Desirability Questioning and the Validity of Indirect Questioning.”

Journal of Consumer Research 20 (2): 303-315.

Fisher, Robert J., and Gerard J. Tellis. 1998. “Removing Social Desirability Bias With Indirect

Questioning: Is the Cure Worse Than The Disease.” *Advances in Consumer Research* 25: 563-567.

Food Standards Agency. 2016. “Guide to creating a front of pack (FoP) nutrition label for

prepacked products sold through retail outlets.” Available at

https://www.food.gov.uk/sites/default/files/media/document/fop-guidance_0.pdf (accessed November 12, 2019).

Gneezy, Uri, and A. Imas. 2017. “Lab in the field: Measuring preferences in the wild.” In

Handbook of economic field experiments, ed. Abhijit Vinajak Banerjee and Esther Duflo, Vol. 1: 439–464. Amsterdam, The Netherlands: North Holland.

Haghani, Milad, and Majid Sarvi. 2019. “Laboratory experimentation and simulation of discrete

direction choices: Investigating hypothetical bias, decision-rule effect and external validity based on aggregate prediction measures.” *Transportation Research Part A: Policy and Practice* 130: 134-157.

Haire, Mason. 1950. “Projective Techniques in Market Research.” *Journal of Marketing* 14 (5):

649-656.

Hanley, Nick, Douglas MacMillan, Robert E. Wright, Craig Bullock, Ian Simpson, Dave Parsisson,

and Bob Crabtree. 1998. “Contingent Valuation Versus Choice Experiments: Estimating the Benefits of Environmentally Sensitive Areas in Scotland.” *Journal of Agricultural Economics* 49 (1): 1-15.

- Harrison, Glenn W. 2014. "Real choices and hypothetical choices." Chapters, in *Handbook of Choice Modelling*, ed. Stephane Hess, and Andrew Daly, 236-254, Edward Elgar Publishing.
- Harrison, Glenn W., Ronald M. Harstad, and E. Elisabeth Rutström. 2004. "Experimental methods and elicitation of values." *Experimental Economics* 7: 123–140.
- Harrison, Glenn W., and John A. List. 2004. "Field experiments." *Journal of Economic Literature* 42 (4): 1009- 1055.
- Harrison, Glenn W., and E. Elisabeth Rutström. 2008. "Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods." In *Handbook of Experimental Economics Results*, ed. Charles R. Plott, and Vernon L. Smith, 752-767, Elsevier
- Harrison, Glenn W., Jimmy Martínez-Correa, J. Tod Swarthout, and Eric R. Ulm. 2017. "Scoring rules for subjective probability distributions." *Journal of Economic Behavior and Organization* 134: 430–448.
- Hensher, David A., and William H., Greene. 2003. "The Mixed Logit model: The state of practice." *Transportation* 30: 133–176.
- Hess, Stephan, David A. Hensher, and Andrew Daly. 2012. "Not bored yet – Revisiting respondent fatigue in stated choice experiments". *Transportation Research Part A: Policy and Practice* 46 (3): 626-644.
- Hole, Arne Risa, and Julie Riise, Kolstad. 2012. "Mixed logit estimation of willingness to pay distributions: a comparison of models in preference and WTP space using data from a health-related choice experiment." *Empirical Economics* 42: 445-469.
- Holmes, Thomas P., Wiktor L. Adamowicz, and Fredrik Carlsson. 2017. "Choice experiments." In *A primer on nonmarket valuation*, ed. Patricia A. Champ, Kevin J. Boyle, and Thomas C. Brown, Ch. 5. Springer, New York.
- Howard, Gregory, Brian E. Roe, Erik C. Nisbet, and Jay F. Martin. 2017. "Hypothetical bias mitigation techniques in choice experiments: do cheap talk and honesty priming effects fade

with repeated choices?” *Journal of the Association of Environmental and Resource Economists* 4 (2): 543-573.

Jacquemet, Nicolas, Robert-Vincent Joule, Stephane Luchini, and Jason F. Shogren. 2013.

“Preference Elicitation under Oath.” *Journal of Environmental Economics and Management* 65: 110-132.

Jacquemet, Nicolas, Alexander James, Stepahne Luchini, and Jason F. Shogren. 2017. “Referenda under Oath.” *Environmental and Resource Economics* 65 (3): 479-504.

Janssen, Ellen M., Deborah A. Marshall, A. Brett Hauber, and John F. Bridges. 2017. “Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability?” *Expert Review of Pharmacoeconomics & Outcomes Research* 17 (6): 531-542.

Johnston, Robert J., Kevin J. Boyle, Wiktor Adamowicz, Jeff Bennett, Roy Brouwer, et al. 2017.

“Contemporary Guidance for Stated Preference Studies.” *Journal of the Association of Environmental and Resource Economists* 4 (2): 319 – 405.

Kealy, Mari Jo, Mark Montgomery, and John F. Dovidio. 1990. “Reliability and Predictive Validity of Contingent Values: Does the Nature of the Good Matter?” *Journal of Environmental Economics and Management* 19: 244-263.

Kemper, Nathan P., Jennie S. Popp, and Rodolfo M. Nayga Jr. 2020. “A query theory account of a discrete choice experiment under oath.” *European Review of Agricultural Economics* 47 (3): 1133–1172.

Krinsky, Itzhak, and A. Leslie Robb. 1986. “On approximating the statistical properties of elasticities.” *The Review of Economics and Statistics* 68 (4): 715-719.

Lewis, Karen E., Carola Grebitus, and Rodolfo M. Nayga Jr. 2016. “U.S. consumers’ preferences for imported and genetically modified sugar: Examining policy consequentiality in a choice experiment.” *Journal of Behavioral and Experimental Economics* 65: 1-8.

- Liebe, Ulf, Jürgen Meyerhoff, and Volkmar Hartje. 2012. "Test-retest reliability of choice experiments in environmental valuation." *Environmental and Resource Economics* 53 (3): 389–407.
- Lin, Wen, David L. Ortega, and Vincenzina Caputo. 2019. "Are Ex-Ante Hypothetical Bias Calibration Methods Context Dependent? Evidence from Online Food Shoppers in China." *The Journal of Consumer Affairs* 53 (2): 504-544.
- List, John A., and Craig A., Gallet. 2001. "What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values?" *Environmental and Resource Economics* 20 (3): 241-54.
- Lloyd-Smith, Patrick, Ewa Zawojka, and Wiktor Adamowicz, 2021. "Moving beyond the Contingent Valuation versus Choice Experiment Debate: Presentation Effects in Stated Preference." *Land Economics* 96 (1): 1–24.
- Loewenstein, George 1999. "Experimental Economics from the Vantage-Point of Behavioural Economics." *The Economic Journal* 109 (453): 25-34.
- Loomis, John, 2011. "What's To Know About Hypothetical Bias In Stated Preference Valuation Studies." *Journal of Economic Surveys* 25 (2): 363-370.
- Loomis, John, 2014. "WAEA Keynote Address: Strategies for Overcoming Hypothetical Bias in Stated Preference Surveys." *Journal of Agricultural and Resource Economics* 39 (1), 34-46.
- Lusk, Jayson L., and F. Bailey Norwood. 2009a. "An Inferred Valuation Method." *Land Economics* 85 (3): 500-514.
- Lusk, Jayson L., and F. Bailey Norwood. 2009b. "Bridging the gap between laboratory experiments and naturally occurring markets: An inferred valuation method." *Journal of Environmental Economics and Management* 58: 236-250.
- Macdiarmid, Jennie I., Simone Cerroni, Dimitrios Kalentakis, and Christian Reynolds. 2021. "How Important is Healthiness, Carbon Footprint and Meat Content When Purchasing a Ready

Meal? Evidence from a Non-Hypothetical Discrete Choice Experiment.” *Journal of Cleaner Production* 282: 124510.

Mamkhezri, Jamal, Jennifer A. Thacher, Janie M. Chermak, and Robert P. Berrens. 2020. “Does the solemn oath lower WTP responses in a discrete choice experiment application to solar energy?” *Journal of Environmental Economics and Policy* 9 (4): 447-473.

Mariel Petr, David Hoyos, Jürgen Meyerhoff, Mikolaj Czajkowski, Thijs Dekker et al. 2021. “Validity and Reliability.” In *Environmental Valuation with Discrete Choice Experiments*, ed. Mariel Petr, David Hoyos, Jürgen Meyerhoff, Mikolaj Czajkowski, Thijs Dekker et al., 111-123. SpringerBriefs in Economics. Springer, Cham.

Matthews, Yvonne, Riccardo Scarpa, and Dan Marsh. 2017. “Stability of willingness-to-pay for coastal management: a choice experiment across three time periods”. *Ecological Economics* 138: 64–73.

McFadden, Daniel 1973. “Conditional Logit Analysis of Qualitative Choice Behaviour.” In *Frontiers in Econometrics*, ed. P. Zarembka, 105-142. Academic Press, New York.

McFadden, Daniel and Kenneth Train. 2017. “*Contingent Valuation of Environmental Goods*.” Northampton, MA: Edward Elgar.

McNair, Ben J., Jeff Bennett, and David A. Hensher. 2011. “A comparison of responses to single and repeated discrete choice questions.” *Resource and Energy Economics* 33 (3): 554-571.

Menapace, Luisa, and Roberta Raffaelli. 2020. “Unraveling hypothetical bias in discrete choice experiments.” *Journal of Economic Behavior & Organization* 176: 416-430.

Mitani, Yohei, and Nicholas Flores. 2014. “Hypothetical bias reconsidered: Payment and provision uncertainties in a threshold provision mechanism.” *Environmental & Resource Economics* 59 (3): 433-454.

Mitchell, Robert Cameron, and Richard T Carson. 1989. “Using surveys to value public goods: the contingent valuation method.” *Resources for the Future*, Washington, DC

- Mørkbak, Morten Raun, and Søren Bøye Olsen. 2014. "A within-sample investigation of test–retest reliability in choice experiment surveys with real economic incentives." *Australian Journal of Agricultural and Resource Economics* 59: 375–392.
- Murphy, Allan H., and Robert L. Winkler. 1970. "Scoring Rules in Probability Assessment and Evaluation." *Acta Psychologica* 34: 273–286.
- Murphy, James, P. Geoffrey Allen, Thomas Stevens, and Darryl Weatherhead, 2005. "A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation." *Environmental & Resource Economics* 30 (3): 313–325.
- Norwood, F. Bailey, and Jayson L. 2011. "Social Desirability Bias in Real, Hypothetical, and Inferred Valuation Experiments." *American Journal of Agricultural Economics* 93 (2): 528–534.
- Oehlmann Malte, and Jürgen Meyerhoff. 2017. "Stated preferences towards renewable energy alternatives in Germany – do the consequentiality of the survey and trust in institutions matter?" *Journal of Environmental Economics and Policy* 6 (1): 1–16.
- Özdemir, Semra, F. Reed Johnson, and A. Brett Hauber. 2009. "Hypothetical bias, cheap talk, and stated willingness to pay for health care." *Journal of Health Economics* 28 (4): 894–901.
- Penn, Jerrod, and Wuyang Hu. 2018. "Understanding Hypothetical Bias: An Enhanced Meta-Analysis." *American Journal of Agricultural Economics* 100 (4): 1186–1206.
- Poe, Gregory L., Kelly L., Giraud, and John B. Loomis. 2005. "Computational methods for measuring the difference of empirical distributions." *American Journal of Agricultural Economics* 87 (2): 353–365.
- Prelec, Drazen 2004. "A Bayesian Truth Serum for Subjective Data." *Science* 306: 462–466.
- Price, J., D. Dupont, and Wiktor Adamowicz. 2017. "As Time Goes By: Examination of Temporal Stability Across Stated Preference Question Formats." *Environmental & Resource Economics* 68: 643–662.

- Radanovic, Goran, and Boi Faltings. 2014. "Incentives for truthful information elicitation of continuous signals." *Proceedings of the 28th AAAI Conference on Artificial Intelligence* (AAAI14).
- Raffaelli, Roberta, Mariangela Franch, Luisa Menapace, and Simone Cerroni. 2021. "Are tourists willing to pay for decarbonizing tourism? Two applications of indirect questioning in discrete choice experiments." *Journal of Environmental Planning and Management*, <https://doi.org/10.1080/09640568.2021.1918651>
- Rakotonarivo, O. Sarobidy, Marije Schaafsma, and Neal Hockley. 2016. "A systematic review of the reliability and validity of discrete choice experiments in valuing non-market environmental goods." *Journal of Environmental Management* 183: 98-109.
- Rigby, Dan, Michael Burton, and Jo Pluske. 2016. "Preference Stability and Choice Consistency in Discrete Choice Experiments." *Environmental & Resource Economics* 65: 441–461.
- Ryan, Mandy, Ann Netten, Diane Skåtun, and Paul Smith. 2006. "Using discrete choice experiments to estimate a preference-based measure of outcome - An application to social care for older people." *Journal of Health Economics* 25: 927–944.
- Ryan, Mandy, and Verity Watson. 2009. "Comparing welfare estimates from payment card contingent valuation and discrete choice experiments." *Health Economics* 18: 389-401.
- Schaafsma, Marije, Roy Brouwer, Inge Liekens, and Leo De Nocker. 2014. "Temporal stability of preferences and willingness to pay for natural areas in choice experiments: a test-retest." *Resource and Energy Economics* 38: 243-260.
- Shogren, Jason F. 2005. "Experimental Methods and Valuation". In *Handbook of Environmental Economics*, ed. Karl Göran Mäler, Jeffrey R. Vincent, 969-1027, Elsevier.
- Shogren, Jason F. 2006. "Valuation in the lab." *Environmental & Resource Economics* 34: 163–172.

- Shogren Jason F. 2010. "Experimental methods in environmental economics." *Behavioural and Experimental Economics. The New Palgrave Economics Collection*, ed. Steven N. Durlauf, and Lawren E. Blume, 137-145. Palgrave Macmillan, London.
- Silva, Andres, Rodolfo M. Nayga, Jr., Benjamin Campbell, and John L. Park. 2011. "Revisiting Cheap Talk with New Evidence from a Field Experiment." *Journal of Agricultural and Resource Economics* 36: 280-291.
- Skjoldborg, Ulla Slothuus, Jørgen Lauridsen, and Peter Junker. 2009. "Reliability of the Discrete Choice Experiment at the Input and Output Level in Patients with Rheumatoid Arthritis." *Value in Health* 12 (1): 153-158.
- Smith, Vernon L. 1976). "Experimental economics: Induced value theory." *American Economic Review* 66: 274–279.
- Thiene, Mara, and Riccardo Scarpa. 2009. "Deriving and Testing Efficient Estimates of WTP Distributions in Destination Choice Models." *Environmental & Resource Economics* 44: 379-395.
- Train, Kenneth, and Melvyn Weeks. 2005. "Discrete choice models in preference space and willingness-to- pay space." In *Applications of Simulation Methods in Environmental and Resource Economics*, ed. Riccardo Scarpa , Anna Alberini, 1-16. Dordrecht, The Netherlands: Springer Publisher.
- UK Government, 2019. *National Minimum Wage and National Living Wage rates* <https://www.gov.uk/national-minimum-wage-rates>. Accessed 17/06/2021.
- van der Pol, Marjon, Alan Shiell, Flora Au, David Johnston, and Suzanne Tough. 2008. "Convergent validity between a discrete choice experiment and a direct, open-ended method: Comparison of preferred attribute levels and willingness to pay estimates." *Social Science & Medicine* 67 (12): 2043-2050.

- Vossler, Christian A., and Mary F., Evans. 2009. "Bridging the Gap between the Field and the Lab: Environmental Goods, Policy Maker Input, and Consequentiality." *Journal of Environmental Economics and Management* 58: 338-345.
- Vossler, Christian A., Maurice Doyon, and Daniel Rondeau, 2012. "Truth in Consequentiality: Theory and Field Evidence on Discrete Choice Experiments." *American Economic Journal: Microeconomics* 4: 145- 171.
- Vossler, Christian A, and Ewa Zawojkska. 2020. "Behavioral drivers or economic incentives? Toward a better understanding of elicitation effects in stated preference studies." *Journal of the Association of Environmental and Resource Economists* 7 (2): 279-303.
- Witkowski, Jens, and David C. Parkes. 2012. "A Robust Bayesian Truth Serum for Small Populations." *Proceedings of the 26th AAAI Conference on Artificial Intelligence* (AAAI13).
- Wuepper, David, Alexandra Clemm, and Philipp Wree. 2019. "The preference for sustainable coffee and a new approach for dealing with hypothetical bias." *Journal of Economic Behaviour & Organisation* 158: 475-486.
- Yadav, Lava, Thomas M. van Rensburg, and Hugh Kelley. 2013. "A Comparison Between the Conventional Stated Preference Technique and an Inferred Valuation Approach." *Journal of Agricultural Economics* 64 (2): 405-422.
- Yangui, Ahmed, Faical Akaichi, Montserrat Costa-Font, and José Maria Gil. 2019. "Comparing results of ranking conjoint analyses, best–worst scaling and discrete choice experiments in a non-hypothetical context." *Australian Journal of Agricultural and Resource Economics* 63: 221-246.
- Zawojkska, Ewa, Anna Bartczak, and Mikołaj Czajkowski. 2019. "Disentangling the effects of policy and payment consequentiality and risk attitudes on stated preferences." *Journal of Environmental Economics and Management* 93: 63-84.

Zizzo, Daniel John 2010. "Experimenter demand effects in economic experiments." *Experimental Economics* 13 (1): 75-98.

CHOICE SITUATION

	OPTION A: COTTAGE PIE	OPTION B: COTTAGE PIE	OPTION C: NO BUY
SATURATED FAT	<div>1.2g</div>	<div>2.3g</div>	
SALT	<div>2.3g</div>	<div>1.1g</div>	
ANITIOTIC RESIDUES	Raised without antibiotics	Raised with antibiotics	
PRICE	£2.00	£3.00	
Select your preferred alternative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

CHOICE SITUATION			
	OPTION A: COTTAGE PIE	OPTION B: COTTAGE PIE	OPTION C: NO BUY
SATURATED FAT	<div><div>1.2g</div></div>	<div><div>2.3g</div></div>	
SALT	<div><div>2.3g</div></div>	<div><div>1.1g</div></div>	
ANITIOTIC RESIDUES	Raised without antibiotics	Raised with antibiotics	
PRICE	£2.00	£3.00	
How many participants choose OPTION A, B and C?	__5__ (out of 10)	__4__ (out of 10)	__1__ (out of 10)
Check earnings	__£7.90__	__£6.90__	__£3.90__