Transportation Research Part B 89 (2016) 149–167

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Can you ever be certain? Reducing hypothetical bias in stated choice experiments via respondent reported choice certainty

Matthew J. Beck[a,*], Simon Fifer[b], John M. Rose[c]

[a] Institute of Transport and Logistics Studies, The University of Sydney, 378 Abercrombie St, Darlington, NSW 2006, Australia
[b] Community and Patient Preference Research, Suite 112, 29 Kiora Rd, Miranda, NSW 2228, Australia
[c] Institute for Choice, University of South Australia, Level 13, 140 Arthur St, North Sydney, NSW 2060, Australia

ARTICLE INFO

Article history:
Received 11 May 2015
Revised 4 April 2016
Accepted 5 April 2016
Available online 27 April 2016

Keywords:
Hypothetical bias
Stated preference
Certainty calibration
GPS
Insurance
Revealed preference

ABSTRACT

Stated choice experiments are a preeminent method for researchers and practitioners who seek to examine the behavior of consumers. However, the extent to which these experiments can replicate real markets continues to be debated in the literature, with particular reference to the potential for biased estimates as a result of the hypothetical nature of such experiments. In this paper, a first in the transportation literature, we compare stated choice responses to revealed preference behavior and examine three methods proposed in the literature for calibrating choice experiments via reported choice certainty. In doing so we provide evidence that the incorrect calibration of responses can produce stated choice results that are more biased than doing nothing at all, however we show that by jointly estimating choice and choice certainty there is a significant reduction in hypothetical bias such that stated choice responses more directly replicate real behavior.

1. Introduction

Stated choice methods are used extensively to understand preference structures across a broad range of literatures, in particular to forecast behavioral responses to product or policy changes or to provide estimates of how respondents value both market and non-market attributes of products or services. The widespread adoption of stated choice methods in fields such as marketing, health, environment and transport are due to the advantages this approach offers with respect to reliable estimates of the relative importance of choice attributes, the ability to incorporate attributes that do not currently exist in the market, the ability to minimize confounding between estimates of effects and creation of designs that enable the efficient recovery of such effects. While these benefits exist, there have been a number of criticisms of stated choice surveys, one of which is the potential for hypothetical bias.

1.1. What is hypothetical bias?

One criticism that is often raised of stated choice experiments (and stated preference methods more widely) is that the intentions which are stated in these experiments are not the behaviors which are observed (or revealed) in actual markets (see for example Samuelson, 1955; Cummings et al., 1986; Mitchell and Carson, 1989). This discrepancy is broadly termed "hypothetical bias" in the literature, as it is argued that bias is generated by the hypothetical nature of a stated preference

* Corresponding author. Tel.: +61 2 9114 1834.
E-mail addresses: matthew.beck@sydney.edu.au (M.J. Beck), simon.fifer@cappre.com.au (S. Fifer), john.rose@unisa.edu.au (J.M. Rose).

http://dx.doi.org/10.1016/j.trb.2016.04.004
0191-2615/

experiment wherein respondents are not obliged to actually carry out the choices or behaviors they state, or are unable to fully predict their own real market behavior in a hypothetical setting like a stated choice survey. There is no widely accepted general theory of respondent behavior that explains hypothetical bias (Loomis, 2011), so in the context of this paper we use the term hypothetical bias to simply refer to discrepancies between preferences exhibited in the hypothetical choice experiment and the preferences revealed by actual behavior.

Before addressing hypothetical bias in more detail, we would like to draw the reader's attention to a wider concept. In much, if not all, of the literature on hypothetical bias there is very little reference to external validity. This connection should be made as external validity has been a long standing concept in statistics. External validity relates chiefly to generalizability; the degree to which the conclusions in a study would hold for other persons in other places at other times. For a comprehensive review of external validity in the social sciences we refer to the reader to Lucas (2003). Included in this paper is a summary of the typical approaches to which external validity is assessed (many of which are applicable to studies of stated choice). Specifically, to have external validity a study must ensure: *construct validity* (the extent to which measure accurately reflect the theoretical concepts they are intended to measure); *relevance* (the degree to which the experimental situation designed to capture the theory adheres to the scope of the theory being tested); *reproducibility* (research findings can be replicated successfully if the research is performed again); *consistency* (the extent to which observations in the study are consistent with each other and the theory being tested); and *confirmatory* (the extent to which the proposition has been supported in numerous tests in diverse settings.

There has been work in the stated preference literature that looks broadly at external validity, typically making combined use of stated and revealed preference data to assess how responses from stated preference experiments can be generalized to real market behavior (see for example Louivere, 1988; Ben-Akiva et al., 1994 and Herriges et al., 1999) using methods such as the nested-logit or error components model to ensure parameter estimates across data types are not confounded by differences in scale (Hensher and Bradley, 1993; Hensher et al., 2008). These methods (and those used in the rest of this paper) are focused on aligning stated preference data with revealed behaviors, which is only partly concerned with the concept of reproducibility; partly because all of the studies on hypothetical bias examine methods to bring the results from one study (on one sample) of stated preferences to one study (on one sample) of revealed behaviors. It is our belief that hypothetical bias, in the literature, has come to mean external validity, whereas external validity is a much more exhaustive concept that requires a number of conditions to be satisfied.

While our discussion addresses methods to reduce hypothetical bias in stated preference studies, an area to which much more attention should be given, we feel that it is important to state that reductions in hypothetical bias do not automatically equate to external validity, rather it is a narrow focus on a much wider issue.

1.2. Does hypothetical bias exist?

The exploration of this source of bias has been most extensive within the contingent valuation literature. A 1994 issue of the Journal of Economic Perspectives included a symposium on contingent valuation, with three articles discussing the role of hypothetical bias in this method (Portney, 1994; Hanemann, 1994; Diamond and Hausman, 1994). For a broader overview of the impact of hypothetical bias in contingent valuation methods, List and Gallet (2001), Little and Berrens (2004) and Murphy et al., (2005) all perform meta-analysis on a number of contingent valuation studies and conclude that hypothetical bias is a major concern, with median bias levels ranging anywhere from 25% to 300%.

Given the significant differences between the stated and revealed values in contingent valuation studies, it is unsurprising that similar exploration has begun to emerge in the stated preference literatures. Evidence of hypothetical bias can be observed in a transportation context, for example Brownstone and Small, (2005) observed that, in a toll road study in California, the revealed value of travel time savings in the morning commute is $20–$40 per hour, which is more than double the values estimated from stated preference studies of the same travel choices. This evidence is supported by Isacsson, (2007) who finds that stated preference values of time are understated for buses, and by Wardman and Shires, (2001) who find that stated valuations similarly overestimate actual values in the context of penalties for having to change trains. In contrast, Wardman and Whelan, (2001) find implausibly large stated values for new or improved trains across 45 stated preference studies. Hensher, (2010) compares values of travel time for a number of different data sets and concludes that differences are generally not statistically significant. Loomis, (2011) also found evidence that hypothetical willingness to pay values exceed actual values by a factor of two to three, though is not always present in stated choice surveys.

Outside of transportation, three of the more interesting studies that examine the phenomenon of hypothetical bias can be found. Chang et al., (2009) compare hypothetical choices of ground beef, wheat flour and dishwashing liquid to actual retail shopping behavior and find evidence that hypothetical choices are a poor predictor of changes to market share, as calculated by the mean square error around predicted and actual shares in each product category. Miller et al., (2011) examine choice of a cleaning product for high-tech equipment, using a tailored online store to gather real purchasing data, and also find the existence of hypothetical bias. Interestingly however, they conclude that stated preference experiments may still lead to the right demand curves and right pricing decisions. Hudsen et al., (2012) investigate the choice of a new product, freshwater prawns, using a mail survey and a controlled in-store experiment. Overall the authors find that hypothetical bias is not present in the choice of the new product (freshwater prawns) but that it is present in the choice of the substitute product (lobster).

In an interesting paper, Lancsar and Swait, (2014) also call for greater attention to the external validity of choice experiments, arguing that investigation of external validity extends beyond the analysis of final outcomes or predictive accuracy of models, but should examine process validity; the extent to which the actions and thought processes of test takers or survey responders demonstrate that they understand the construct in the same way it is defined by the researchers. They argue that process validity is broader, incorporating the choice context, choice process, task design and econometric models that more accurately capture the choice reality faced by respondents.

The aforementioned studies highlight the complex nature of hypothetical bias. While the usual assumption, particularly for public goods, is that hypothetical bias leads to overstated valuations compared to true values (Champ et al., 1997, Carson and Groves, 2007, Harrison and Rutstrom, 2008), it is equally possible that it manifests itself as understated values, particularly in the context of private goods (Loomis et al., 2000, Harrison et al., 2004, Harrison, 2006). What is evident from the literature discussed above is that there is no clear a priori assumption as to how hypothetical bias may manifest itself, if at all. Given that hypothetical bias does not express itself in a predictable way, there is an unsurprisingly wide range of varying approaches in the extant literature that seek to mitigate it. These methods are applied either ex-ante or ex-post or some combination of both (Whitehead and Cherry, 2007).

1.3. Methods of mitigating hypothetical bias

One of the main arguments as to why hypothetical bias exists is that, given there is no direct payoff or penalty associated with the choices made in a hypothetical environment, stated preference methods lack salience. Without any real economic commitment, the consumer is free to behave differently to the way they would if they were required to replicate those decisions in a real market (Harrison, 2007, Hensher, 2010). In a typical stated preference experiment there is no directive incentive for respondents to reveal their true preferences; in the instances where respondents do receive an incentive it is for participation and not directly linked to the quality of responses provided. Consequently, there has been research into methods that seek to make such experiments incentive compatible which, based on the seminal work of Wilson, (1978) and Myerson, (1979), is defined as a process whereby the respondents are better off truthfully revealing their private information as asked for by the process. The majority of ex-ante methods are designed to mitigate this potential source of hypothetical bias.

Carson and Groves, (2007) describe the revealing of true preferences as an outcome of whether the participant cares about the results of the research, and believes that his or her answers will influence the decisions what are made as a result of the research; that the survey responses must be consequential to the respondent. Typically such studies involve the administration of a standard stated choice experiment except respondents are informed upfront that, once their choices are made, one choice will be selected at random to be binding (see for example; Lusk and Schroeder, 2004; Ding et al., 2005; List et al., 2006; Alfnes et al., 2006; Chowdhury et al., 2010; Carlson et al., 2010; Miller et al., 2011; and Moser et al., 2010). The current evidence for incentive alignment suggests that it may be a useful method for reducing the hypothetical bias associated with stated preference experiments. However, this method is primarily relevant for low-involvement consumer-driven private goods, because of the binding purchase requirement. For larger scale purchases such as motor vehicles, houses or holidays or with lifestyle ramifications such as health care or insurance, the creation of a binding constraint is impossible.

An alternative method used to incentivize respondents to reveal their true preferences is to make it clear to respondents that the results of the survey will be used by decision makers in ways that will affect goods or services made available to the respondent (Vossler and Evans, 2009, Herriges et al., 2010, Vossler et al., 2012). Akin to this method is an approach termed "cheap talk" where the respondent is explicitly told of the existence of hypothetical bias and emphasizes the importance of the study to respondents prior to completing the choice task. This method has had varying degrees of success (positive experiences include Cummings and Taylor, 1999, List, 2001, Aadland and Caplan, 2003, Landry and List, 2007 and negative experiences include Aadland and Caplan, 2006, Blumenschein et al., 2008). A more extreme version of this approach actually requires respondents to sign a document promising they will tell the truth (Jacquemet et al., 2013, Stevens et al., 2013).

While ex-ante methods focus on encouraging respondents to act more realistically, the ex-post methods focus on refining the data that is collected. General approaches include removing abnormal observations (willingness to pay values that are too large or too small) and re-estimating models on the cleaned data (Mitchell and Carson, 1989, Haab and McConnell, 2002, Davies and Loomis, 2010). Alternatively, the stated preference models can be recalibrated against real market results to replicate market shares (Fox et al., 1998, Hensher, 2010). However, one method that is showing clear promise as a method to better align stated and actual willingness to pay values is certainty calibration (Loomis, 2014), a method of correction that exhibits procedural invariance (Brouwer et al., 2010).

1.4. Respondent reported certainty calibration

Certainty calibration involves the use of a question that asks respondents how certain they are about the choices they make in the stated preference experiment, in particular how certain they are that they would make the same choice in a real market. The underlying behavioral premise behind this method is that respondents may make inconsistent choices in a hypothetical market because they have not formed preferences about how they value the good or service or they may be unfamiliar with the choice object entirely. In collecting data about choice certainty, a question is usually asked after each choice task and framed on a scale of 1 (very uncertain) to 10 (very certain). Other variants on this approach also exist, such

as asking respondents to express certainty as a percentage or on a qualitative scale such as “probably sure” or “definitely sure”.

Responses that exhibit too much uncertainty are removed from estimation such that willingness to pay values are based only on choices that respondents are sure they would repeat in a real context. The threshold at which to separate certain choices from uncertain choices has been found to vary from study to study. For example, Ethier et al., (2000) use responses greater than seven, Champ and Bishop, (2001) use only responses greater than eight, and Champ et al., (1997) restrict their analysis to responses that achieved a certainty score of ten. It is also suggested that analysts report a variety of estimates at differing cut-off points (Champ et al., 2009). Despite the variations in how certainty is employed, the use of such a technique has proven to be useful in eliminating potential bias induced by the hypothetical nature of an experiment (Li and Mattson, 1995, Champ et al., 1997, Johannesson et al., 1999, Ethier et al., 2000, Blumenschein et al., 2008, Champ et al., 2009; Morrison and Brown, 2009; Moore et al., 2010).

While the use of certainty calibration in contingent valuation has been extensive, only tentative steps have been taken to examine the role such a technique may play in choice experiments. Olsson, (2005), in perhaps the first study of certainty calibration in choice experiments, found that willingness to pay measures are positively and significantly related to degree of response uncertainty, such that more certain respondents have higher willingness to pay. Promisingly, in examining donation behavior, Norwood, (2005) found that eliminating responses less than eight (on a ten point scale) bought the results from the hypothetical scenarios into alignment with those from the real donation scenarios. Similar results were also found by Ready et al., (2010), where certainty calibration was successful in achieving results which were similar to real donations for an animal protection program. Beck et al., (2013) examine a range of different approaches to incorporating uncertainty into model estimation and find that results from the modeling process differ across methods, concluding that more work on the use of certainty calibration in choice experiments is required.

While ostensibly introduced to mitigate hypothetical bias, it can also be used to capture other effects within stated preference experiments. Another source of uncertain responses is how complex respondents find the choice task (Swait and Adamowicz, 2001, DeShazo and Fermo, 2002, Arentze et al., 2003, Caussade et al., 2005). Alternatively, research has found that respondents suffer a degree of fatigue in answering choice tasks, with later responses being more inconsistent or uncertain than those made in the initial tasks (Bradley and Daly, 1994). On the other hand, evidence also exists that respondents learn about their preferences as they complete multiple choice tasks and thus the later choices exhibit a greater degree of certainty and consistency (Brazell and Louviere, 1998). It has also been found that while self-reported certainty indicates that respondents feel that learning occurs, there is no econometric evidence for either learning or fatigue effects (Brouwer et al., 2010). Overall, certainty calibration has been used extensively and to positive effect in contingent valuations studies and evidence from such calibration in choice experiments provides early promise.

1.5. Contribution of this study

While the literature exploring certainty calibration as a method for accounting for hypothetical bias in choice experiments identifies it has having promise and improving fit (for example Brouwer et al., 2010, Beck et al., 2013), to the best of our knowledge there exists only two studies that examine how well certainty indexing performs as calibration method and both are in the context of donation behavior (a public good). Specifically, Norwood, (2005) who constructed an experiment around students donating grade points to a common pool to be reallocated to all students and Ready et al., (2010) who examined if certainty calibration was successful in achieving results which were similar to real donations for an animal protection program.

Similar to both these studies, we make use of an innovative revealed preference experiment to assess how effective different methods of certainty calibration are in eliminating or reducing hypothetical bias, however we differ in that we examine choice behavior in the context of a private good and we compare multiple methods of certainty calibration. Specifically, we compare the most common method of certainty calibration (which is to remove or recode choices where uncertainty is deemed to be too high), to a method proposed in Beck et al., (2013) where choice tasks are probability weighted based on the level of uncertainty expressed (an approach that produced the greatest improvements in model fit relative to a range of other approaches assessed in that paper) and to a final method proposed by Rose et al., (2015) which jointly models choice and choice certainty as a method to eliminate potential for endogeniety bias between these two constructs.[1]

Given the lack of clarity around the usefulness of certainty indexing in stated choice experiments as well as conjecture over the appropriate framework for such calibration (Beck et al., 2013), this paper will further develop the theory of certainty calibration within stated choice experiments by exploring a range of metrics via which the certainty scale may be introduced to the estimation of results from hypothetical scenarios, as a means of better replicating the sensitivities observed from real world choices.

[1] One general criticism that can be levelled at previous studies of certainty calibration is that they treat certainty as an exogenous construct, thus failing to take into account that self-reported certainty could potentially be correlated with the unobserved component of utility, meaning that the potential for endogeneity bias is created. In seeking to overcome this criticism, Rose et al. (2015) provide a framework for the simultaneous estimation of certainty and choice, thus eliminating endogeneity bias. Dekker et al. (2014) provide an alternative approach, proposing the analyst treat certainty as a latent construct in the utility function as the method of overcoming endogeneity.

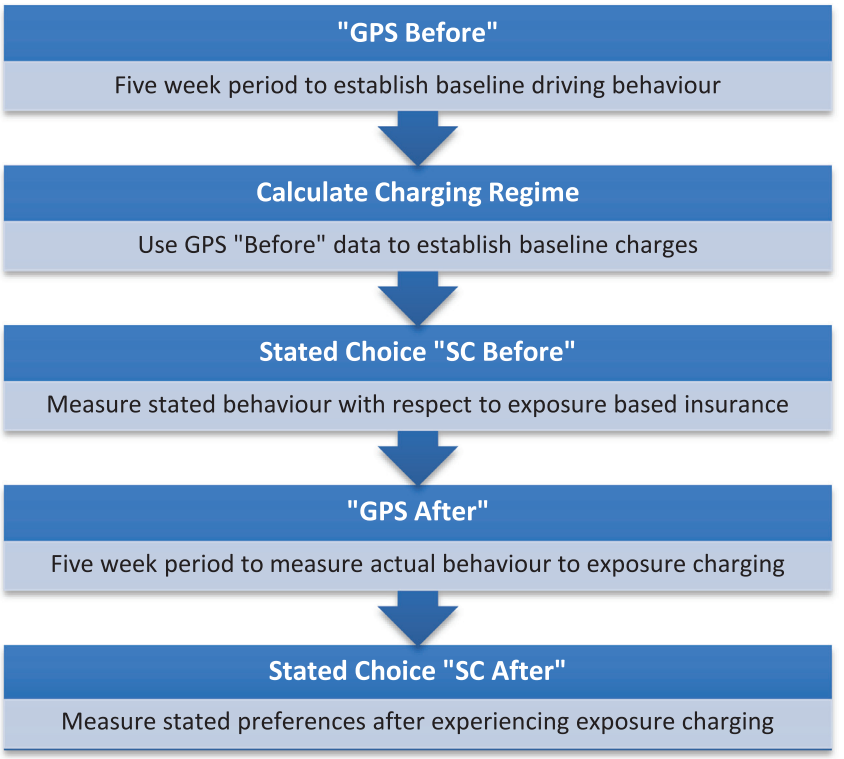

Fig. 1. Overview of the Experiment.

The rest of this paper is structured as follows: In the next section an overview of the empirical context and data characteristics is given. Section 3 provides a review of the methodology used to establish hypothetical bias. Section 4 presents the results of the empirical modeling and discusses the prevalence of hypothetical bias in this study. Finally, Section 5 provides discussion and concluding remarks, highlighting directions of future research.

2. Empirical context and survey design

2.1. Project overview

Within Australia, vehicle insurance is purchased as an annual premium that is typically calculated based on your age, gender, accident record, type of vehicle being insured and level of coverage desired. The overarching aim of the study from which this data is generated is to examine the stated and actual response of drivers to an innovative exposure based insurance charge that additionally varies based on where, when and how you drive. The data used was collected from 148 Sydney motorists who were recruited to take part in a 10-week GPS driving field study in the second half of 2009. The study consisted of five distinct phases which are outlined in Fig. 1. The unique structure of this research facilitated comparison of stated versus revealed choice data from the same sample, for the purpose of investigating hypothetical bias.

The initial “GPS Before” stage involved a period of GPS monitoring during the first five weeks of the study. This period was designed to measure the baseline driving behavior of participants prior to the commencement of the insurance charging regime. Each driver’s baseline behavior also fed into the second stage of the experiment in order to establish a charging regime that would sufficiently incentivize each driver to potentially change their driving habits, by charging motorists according to known correlates of crash risk (age, kilometers driven, night-time driving, and speeding). The charging scheme consisted of a base kilometer charge for the two age groups, younger drivers (17–30 years) and older drivers (31–65 years), with multipliers applied to the base rate charges for driving at night and speeding (for a detailed discussion of how the charging regime was established refer to Fifer et al., 2011).

After initial driving behaviors were observed and relevant exposure based charges were established, approximately half of the participating sample were randomly selected to complete a stated choice survey. The “SC Before” stage was designed to capture what these participants indicated they would do as opposed to what they actually did in reaction to an exposure-based insurance product. The entire sample then participated in a five week “GPS After” phase, where exposure based charging regime was implemented for each participant. A web-based interface enable participants to log on and how they were being charged based on their observed driving patterns and could then use this information to adjust their behavior.

Table 1
Attributes and levels in stated choice experiment.

Attribute	Description	Pivot Levels (off the reference driving behaviour)
Distance	The total number of KM you drive. The number of travel days driving for that purpose is also shown	−75%, −30%, −15%, 0%
Driving time of day	The percentage of your total driving in the 'Day' (5am-8am) and 'Night' (8pm to 5am)	−100%, −75%, −50%, −25%, 0%
Speeding	The percentage of your total driving where you are 'Speeding' and 'Not Speeding'	−75%, −50%, −25%, 0%
Travel time	The average increase in travel time per trip (in minutes) if you were to reduce your speeding behavior	0 min, 2 min, 4 min, 6 min, 8min
Charges	The amount of money you would pay (reduced from your base incentive) to drive for that trip purpose	−75%, −50%, −50%, −20%, -10%

Following the five week "GPS After" period approximately half the sample was randomly selected to complete a stated choice survey as part of the "SC After" phase. This data were collected to model the preferences of participants who had had an experience with the charging regime. The "SC After" subsample consisted of participants who had been randomly assigned to this wave at the beginning of the study and also a repeat group of participants who had already completed the "SC Before" survey. For the purposes of this study, the data used for analysis herein is from the "SC Before" and "GPS After" phases of the study.

2.2. Details of the choice experiment

The purpose of the choice experiment was to explore how respondents might hypothetically change their driving behavior if they were to purchase an exposure-based insurance charging product (Fifer et al., 2011). The survey was designed to correspond closely to the RP decision context to enable a valid examination of the extent of hypothetical bias. The attributes in the experiment included distance split by trip purpose (work; shopping/personal business; social/recreational), time of day driving was done, speeding as percent of all driving, average travel time increase per trip if speeding reduced and the charge applied for risky driving. The number of travel days over the five week period on which that purpose was driven is also shown. The survey was administered online in two waves to a total of 125 respondents who completed both the full five week "GPS Before" revealed preference study. A description of the attributes is displayed in Table 1.

A Bayesian efficient design for each trip purpose was generated. This experimental design method was used to produce lower standard errors and provide more reliable parameter estimates for a relatively small sample size. Additionally, in keeping with recent literature on referencing stated preference experiments to a known experience, a pivot design was also used (Rose et al. 2008; Rose and Bliemer 2009). Such designs take the existing or currently chosen alternative and pivot around this reference point, making incremental changes to the attributes of this reference alternative. In this experiment the reference alternative is the aggregate driving behaviors observed over the five week "GPS Before" phase for each of the three trip types examined (work, shopping/personal business, and social/recreational). In this baseline behavior it was observed how far respondents drove over the course of the week, how often they drove at day versus at night and what percentage of the time they were speeding. The other two alternatives which comprise each choice task, an example of which is provided in Fig. 2, represent alternatives where the respondent could choose to state that they would engage in driving behavior which would reduce their exposure to risk, reductions in which are based on the percentages described in Table 1.

A potential lump-sum reward for safer driving was generated for each respondent based on their level of exposure observed in the "GPS Before" phase (for a full discussion of how this charge was calculated refer to Greaves and Fifer 2010). The range of base incentives ran from AU$25 to AU$915, with an average of AU$300. Respondents were told this is what would be available to them should they eliminate all risky driving behavior completely, however this incentive payment was reduced if the stated choices made by respondents still exhibited some degree of exposure (with the size of the penalty varying based on the amount of reduction in said exposure). Following each choice task was the certainty index, where respondents stated how certain they were (on a ten point scale) that this choice would also be one that they would make in real life. Respondents only answered choice questions for the trip purposes which they drove. At a maximum, if respondents made all three trips they would be required to complete a total of 12 choice tasks (four for each trip purpose).

3. Methodology

3.1. Generalized mixed logit model

The Generalized Mixed Logit model (GMX) is the chosen methodology for estimation in this paper. In establishing this model form let U_{nsj} denote the utility of alternative j perceived by respondent n in choice situation s. U_{nsj} may be partitioned into two separate components, an observed component of utility, V_{nsj} and a residual unobserved (and un-modeled)

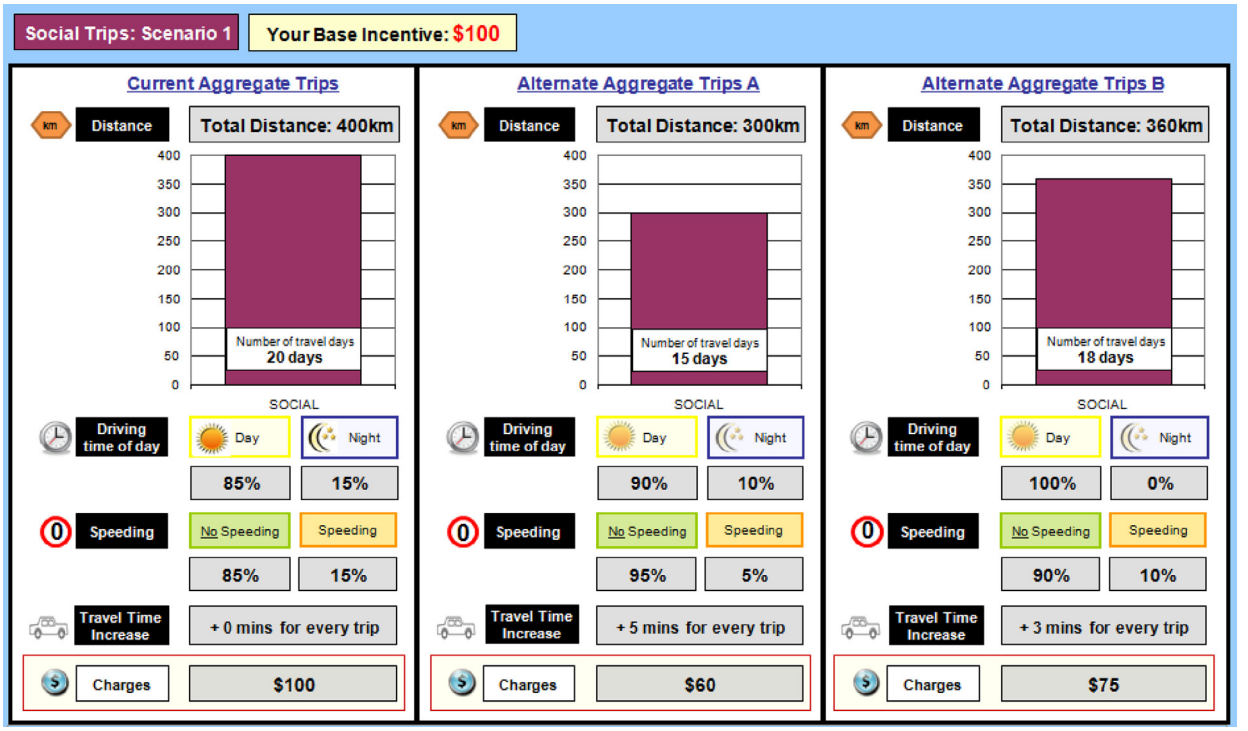

Fig. 2. Example of Stated Choice Task.

component, ε_{nsj}, such that

$$U_{nsj} = V_{nsj} + \varepsilon_{nsj}. \tag{1}$$

The observed component of utility is typically assumed to be a linear relationship of observed attribute levels, x, of each alternative j and their corresponding weights (parameters), β, such that

$$U_{nsj} = \sigma_n \sum_{k=1}^{K} \beta_{nk} x_{nsjk} + \varepsilon_{nsj}, \tag{2}$$

where β_{nk} represents the marginal utility or parameter weight associated with attribute k for respondent n and the unobserved component, ε_{nsj}, is assumed to be independently and identically (IID) extreme value type 1 (EV1) distributed with cumulative distribution function (CDF)

$$F(\varepsilon) = \exp\left(-e^{-\sigma_n(\varepsilon)}\right). \tag{3}$$

This distribution has $E(\varepsilon_{nsj}) = 0.57721/\sigma_n$ and $\mathrm{var}(\varepsilon_{nsj}) = \frac{\pi^2}{6\sigma_n^2}$, where σ_n in Eqs. (2) and (3) represents a positive scale factor that is typically normalized to one in most applications. As such, it can be seen that the variance of ε_{nsj} is inversely related to the magnitude of $\sigma_n \sum_{k=1}^{K} \beta_{nk}$ via σ_n. As such, a preference ordering preserving and alternative representation of Eq. (2) is

$$U_{nsj} = \sum_{k=1}^{K} \beta_{nk} x_{nsjk} + \varepsilon_{nsj}/\sigma_n. \tag{4}$$

As well as containing information on the levels of the attributes, x in Eq. (2) may also contain up to $J-1$ alternative specific constants (ASCs) capturing the residual mean influences of the unobserved effects on choice associated with their respective alternatives; where x takes the value 1 for the alternative under consideration or zero otherwise. The utility specification in Eq. (2) is flexible in that it allows for the possibility that different respondents may have different marginal utilities for each attribute being modeled.

To accommodate both scale and preference heterogeneity, Keane, (2006) proposed the generalized multinomial logit (GMNL) model. First popularized by Fiebig et al., (2010) and subsequently by Greene and Hensher, (2010), the marginal utility for attribute k for the GMNL may be represented as Eq. (5)

$$\beta_{nk} = \sigma_n \bar{\beta}_k + \gamma \eta_k z_n + (1-\gamma)\sigma_n \eta_k z_n, \tag{5}$$

where γ takes any value between 0 and 1 and where

$$\sigma_n = e^{\left(\bar{\sigma}+\sum_{q=1}^{Q}\delta_q w_q+\tau v_n\right)}. \tag{6}$$

In Eq. (6), $\bar{\sigma}$ denotes a mean parameter of scale variance, τ a parameter of unobserved scale heterogeneity and v_n a standard Normal distribution representing the unobserved scale heterogeneity. $\delta_q$in Eq. (6) represents parameters associated with covariates $w_q$which may be used to decompose the scale parameter.

To estimate the GMNL, maximum simulated likelihood estimation is used. In estimating the model however, a number of difficulties must first be overcome. Firstly, both Fiebig et al., (2010) and Greene and Hensher, (2010) note that $\bar{\sigma}$ and σ_n in Eq. (6) cannot be separately identified. Under the assumption that scale heterogeneity will be log-Normally distributed and ignoring preference heterogeneity and ignoring δ_q andw_q, Fiebig et al., (2010) and Greene and Hensher, (2010) both note that

$$E\left[\sigma_n^2\right] = e^{\left(\bar{\sigma}^2+\frac{\tau^2}{2}\right)}. \tag{7}$$

In order for the model to be identified, it is necessary to normalize $E[\sigma_n^2] = 1$.To implement this normalization they set $\bar{\sigma} = \frac{\tau^2}{2}$such that Eq. (7) becomes

$$\sigma_n = \mathrm{e}^{\left(\frac{-\tau^2}{2}+\tau v_n\right)}. \tag{8}$$

Finally, as per Fiebig et al. (2010), γ is restricted to be between zero and one. This restriction occurs by parameterizing γ in terms of α such that

$$\gamma = \frac{e^{\alpha}}{1+e^{\alpha}}, \tag{9}$$

where α is unrestricted in both sign and magnitude.

Under the assumption that the error termε_{nsj}, are IID EV1, the probability that respondent n in choice task s is observed to choose alternative j is given as

$$P_{nsj} = \frac{\exp\left(V_{nsj}\right)}{\sum_{i\in J_{ns}}\exp\left(V_{nsi}\right)}. \tag{10}$$

As the parameters estimates in the GMNL model are assumed to have some continuous density, the probability that respondent n in choice situation s will choice alternative j can be written as

$$L_{nsj} = \int_{\beta} P_{nsj}(\beta) f(\beta|\theta) d\beta, \tag{11}$$

where $f(\beta|\theta)$ is the multivariate probability density function of β, given the distributional parameters θ.

The parameter estimates of the model are located using maximum likelihood methods. Unfortunately, the integral in Eq. (10) does not have a closed analytical form, meaning that it must be evaluated using either Pseudo Monte Carlo or Quasi-Monte Carlo methods, which involves simulating the parameters and choice probabilities, by taking R draws for each of the K random terms or parameters, calculating the choice probabilities for each of the draws., The simulated log-likelihood function is then computed using the expected probability computed from Eq. (11) over the R draws. Assuming independence between the responses of the individual decision makers but not within, the simulated maximum likelihood of the model is given as

$$\begin{aligned} L(E(L_{NS})) &= \log E\left(\prod_{n=1}^{N}\prod_{s\in S_n}\prod_{j\in J_{ns}}\left(P_{nsj}\right)^{y_{nsj}}\right) \\ &= \sum_{n=1}^{N}\log E\left(\prod_{s\in S_n}\prod_{j\in J_{ns}}\left(P_{nsj}\right)^{y_{nsj}}\right). \end{aligned} \tag{12}$$

3.2. Incorporating uncertainty

Each of the four model results presented in this paper are variations of the above model structure. In the stated choice survey, participants answered four choice scenarios across three trip purposes. These were combined for analysis to allow a larger sample size for estimation. In the model, groupings of choice scenario answers for each trip purpose were treated as if they were made by different pseudo-individuals in estimation of the panel effects, to allow for trip purpose differences within individuals. Estimation in this manner accounted for correlation in the preferences of individuals within their set of choice scenarios. In addition, separate constants were estimated for the current alternative to allow for differences in the means of effects that were not observed in the model. The general structure of the model utility functions estimated using the “SC Before” data are:

$$V_{SQ} = \sigma_n \left(ASC_{SQ} + \sum_{k=1}^{K} \beta_{nk} x_{k,SQ} \right)$$
$$V_{AltA} = \sigma_n \left(\sum_{k=1}^{K} \beta_{nk} x_{nk,AltA} \right)$$
$$V_{AltB} = \sigma_n \left(\sum_{k=1}^{K} \beta_{nk} X_{nk,AltB} \right) \quad (13)$$

As discussed previously, these stated choices may be subject to hypothetical bias, so in our attempts to mitigate this we employ three different methods for introducing certainty scaling into the utility function. Specifically, the models estimated in this paper compare the traditional approach of recoding uncertain responses into the status quo, to the weighting approach proposed by Beck et al. (2013) and the joint estimation methodology proposed by Rose et al. (2015). These proposed methods have been shown to produce better model fit statistics relative to other approaches, but they have not been tested against corresponding revealed preference data. Thus, the purpose of the modeling herein is to compare these various approaches to see the improvements in model fit also translate to more accurate measure of willingness to pay statistics.

3.2.1. Recoding the model

The traditional approach to certainly calibration is recoding choices which are deemed as being too uncertain as being either a choice that would not be made, or that the existing behavior would be maintained. This approach is operationalized via data manipulation rather than any change to model estimation. In this study, respondents who chose an alternative that was not the status quo (i.e. Alt A or Alt B as per Eq. (13)), but provided a choice certainty score of less than nine, had that choice recoded back to the choice of the status quo. It should be noted that different thresholds were used to assess certain versus uncertain responses and the threshold certainty scores of nine or more provided the best result (i.e., the lowest level of hypothetical bias observed in the modeling).

3.2.2. Weighting the choices

In the weighting methodology proposed by Beck et al. (2013) (viz. Manski and Lerman 1977), choice tasks are probability weighted via the certainty score assigned to each choice, such that choice tasks with a greater level of certainty are assigned a greater weight and those with lesser respondent certainty are given less weight. In this instance the methodology is operationalized via a simple weighting of the log-likelihood function w_{ns}

$$L(E(L_{NS})) = \sum_{n=1}^{N} w_{ns} \log E \left(\prod_{s \in S_n} \prod_{j \in J_{ns}} \left(P_{nsj} \right)^{y_{nsj}} \right) \quad (14)$$

where w_{ns} is the certainty score given by respondent n in choice task s. Under this formulation choices where the respondent is more certain about that choice are given more weight in estimation than choices where the respondent is uncertain.

3.2.3. Jointly estimating choice and certainty

This method is proposed by Rose et al. (2015) wherein both choice and uncertainty are modeled simultaneously thus eliminating the potential endogeneity between these two responses; a criticism of the use of certainty indexing. This methodology requires the estimation of separate utility functions based on whether the respondent was certain (greater than or equal to nine on the certainty scale) or uncertain (less than nine on the certainty scale), resulting in the specification of six utility functions (Eq. (1) replicated for both certain choices and uncertain choices):

$$V_{SQ,Cert\geq 9} = \sigma_n \left(ASC_{SQ|Cert\geq 9} + \sum_{k=1}^{K} \beta_{nk|Cert\geq 9} x_{k,SQ} \right)$$
$$V_{SQ|Cert<9} = \sigma_n \left(ASC_{SQ|Cert<9} + \sum_{k=1}^{K} \beta_{nk|Cert<9} x_{k,SQ} \right)$$
$$V_{AltA|Cert\geq 9} = \sigma_n \left(\sum_{k=1}^{K} \beta_{nk|Cert\geq 9} x_{nk,AltA} \right)$$
$$V_{AltA|Cert<9} = \sigma_n \left(\sum_{k=1}^{K} \beta_{nk|Cert<9} x_{nk,AltA} \right)$$
$$V_{AltB|Cert\geq 9} = \sigma_n \left(\sum_{k=1}^{K} \beta_{nk|Cert\geq 9} X_{nk,AltB} \right)$$

$$V_{AltB|Cert<9} = \sigma_n \left(\sum_{k=1}^{K} \beta_{nk|Cert<9} X_{nk,AltB} \right) \tag{15}$$

Note that it is possible to have more than one threshold, but in using more thresholds the modeling approach becomes increasingly more complex (Rose et al. 2015). It is important to reiterate that the certainty threshold of nine was used as a result of extensive testing over a range of potential thresholds revealed that the value of nine provided the best result (i.e., the lowest level of hypothetical bias observed in the modeling).

3.2.4. A note on the methodology

In the original literature it was reported that the GMX model accounted for not only preference heterogeneity but also scale heterogeneity between respondents. Despite contrary claims in these early papers, the GMX model fails to identify separate and uncorrelated estimates of scale and preference heterogeneity. Rather, it actually allows for more flexible distributions of heterogeneity through a different parameterization (Hess and Rose 2012).

As can be seen in the previous sections, the methods examined in this paper involve treatments to either the data, or the log-likelihood function or the specification of the utility functions to be estimated. Consequently, the choice of modeling methodology is dependent on the analyst; for example the analyst may choose to deploy a scaled and/or heteroskedastic multinomial logit. We use the GMX as a means of estimating a flexible distributional form. Additionally, other approaches do exist via which certainty scores can be introduced to estimation, for example scale as described by Eq. (4) could be made a function of the certainty scores, however in our trials of models for this paper and in the work of Beck et al. (2013) and Rose et al. (2015) this approach proved less tractable and had inferior model fits.

In this paper, all parameters were treated as random in the two models to allow for estimation of individual conditional parameters. The parameters are called conditional because they are conditioned on the chosen alternatives. Various distributions were tested; however, the constrained triangular distribution provided the best behavioral interpretation. The random parameters for all models were estimated using 2500 Halton draws in Nlogit 5.0.

3.3. Assessing extent of hypothetical bias

Choice experiments are typically used to predict behavior within a market or provide valuation of attributes of a product or service. Consequently, how effective certainty indexing is as a method of better aligning stated choices with revealed choices can be thought of in one of two ways; how well the stated choice data predicts real market behavior or how well stated choice data replicates how attributes are valued in real markets. To account for both applications of stated choice, we measure how much hypothetical bias is reduced via two outputs; differences in total willingness to pay (TWTP) and model predictions.[2]

For each method we report a percentage of respondents affected by hypothetical bias under the various definitions applied and tested herein. This measure was chosen as bias could be exhibited as either an understatement or an overstatement of actual behavior, so an absolute measure of how much hypothetical bias was corrected for was preferred. What constitutes a high prediction value is ambiguous and open to the interpretation of the researcher. To avoid this ambiguity, bias was coded as any incorrect prediction values (i.e., prediction values for the 'After' alternative less than 50%).

3.3.1. Model predictions

The percentage of respondents who exhibit bias in their stated results compared to their actual behavior was determined by first estimating model parameters from the stated choice data. Once these choice equations were estimated, the GPS data collected from the "GPS After" phases were entered into the choice functions (i.e., the actual data collected became the attribute levels used to calculate utilities and thus choice probabilities). The probability of choosing the observed "GPS After" alternative (the actual changes each respondent made in the field study) for each respondent was calculated and this probability was used to define how well the model based on stated choices performs in predicting the actual changes observed in the field study. Throughout this discussion the model probabilities are converted to percentages when referring to model predictions. If the preferences in the choice experiment and the field study are analogous (i.e. there is minimal to low hypothetical bias) we would expect the prediction values for the "GPS After" alternative using the actual data to be high.

3.3.2. Total willingness to pay

In examining hypothetical bias with respect to TWTP we examine the change in consumer surplus from the "GPS Before" and "GPS After" phase. Train (2009) defines the change in consumer surplus as:

$$\Delta E(CS_n) = \frac{1}{-\beta_{Charge}} \left[ln\left(\sum_{j=1}^{J^{After}} e^{V_{nj}^{After}} \right) - ln\left(\sum_{j=1}^{J^{Before}} e^{V_{nj}^{Before}} \right) \right] \tag{16}$$

[2] The approach used to determine how well any form of certainty calibration has done in replicating revealed behavior is dependent on which purpose the stated choice modeling is being done; prediction or valuation. It may also be reasonable to use both statistics as a more holistic assessment of the calibration process.

Table 2
Sample demographics.

Variable	Category	Percent (%)
Gender	Female	53
	Male	47
Age	Old (36–64)	60
	Young (18–35)	40
Household income	Low (<$60k)	32
	Medium (≥$60k ≤$100k)	45
	High (>$100k)	23

The parameters used to calculate the utility from the revealed "GPS After" (V^{After}) and "GPS Before" (V^{Before}) data are those estimated from the "SC Before" phase of the study.

TWTP is calculated for every participant using the mean β parameters of the individual conditional parameter distributions from the choice experiment and the actual data (X's) for the two observed driving periods. Any changes to actual driving behavior between the "GPS Before" and "GPS After" phase are assumed to be a result of the implementation of the exposure based charging scheme.[3]

In this framework, if preferences in reality are the same as preferences expressed in the hypothetical choice task where they were asked to consider their current driving ("GPS Before") and whether they would change it or not in the presence of an exposure based charge, then the utility maximizing behavior is to align real choices with stated choices (i.e. the β's from the stated choice task are the true utility maximizing parameters). Any real behavior that is not aligned with the stated preferences would result in a modeled utility that is suboptimal (i.e. lower than would be predicted by the hypothetical choices). In other words, utility in the "GPS After" period is estimated to be lower than utility from their behavior in the "GPS Before" period, i.e. Eq. (16) is negative. Consequently, any TWTP value that was negative was deemed to be subject to hypothetical bias.

For example, if in the stated choice task with a hypothetical exposure charge a respondent's utility maximizing choice was to change their behavior, if the hypothetical choices were not biased the real behavior would be change their driving behavior. However, if in the real choice the respondent drives exactly the same in the "GPS After" as they did in the "GPS Before" this behavior is not utility maximizing if their hypothetical stated choices are to be believed (i.e. the hypothetical choices are biased). In other words, if the hypothetical exposure charge makes respondents state that their past behavior is no longer utility maximizing, but continue with that past behavior in the presence of a real exposure charge, the predicted utility of this behavior (V^{After}) using those biased parameters would be less than their (V^{Before}) where there was no incentive to change.

4. Results

Sample demographics are provided in Table 2. Note that, as described in Project Overview, drivers were divided into two groups ("old" and "young") in order to determine the relevant exposure charge. It should also be noted that respondents in both the "SC Before" and "GPS After" phase were there same.

4.1. Certainty scale scores

The distribution of certainty scale responses is highlighted in Fig. 3. The majority of respondents were reasonably certain about their stated choice, with an average of 7.41 (with a standard deviation of 1.90) on scale of Very Uncertain (1) to Very Certain (10).

4.2. Model results

The model results shown in Table 3 are for the four models discussed in the methodology: the base model without calibration; the traditional recode method, the probability weighted model; and a jointly estimated certainty / choice model. All parameters are significant and of the expected signs. The use of alternative specific constant allows for the estimation of any differing preference structures specific to that type of trip which are not accounted for by the attributes in the

[3] While changes to driving behaviors may occur due to changes in residential or work location/conditions, for respondents in this study such factors were constant over the duration of the study period. However, other exogenous factors outside the exposure based scheme may also impact on behavior; police blitzes on speeding or unsafe driving might promote behavioral change, being involved in, witnessing or knowing someone involved in a traffic incident may also cause change, as might large changes to leisure choices which in turn affect driving behavior. In the collection of the data respondents were asked if anything major happened that might influence their driving behavior. Respondents who provided such a reason were excluded from analysis. While we acknowledge that we may not be able to completely account for these confounding factors, we believe that any change that could be attributed to such exogenous factors would affect only a small part of the sample, if at all.

Table 3
Generalized mixed logit results.

	Base model		(1) Recode method		(2) Weighting method		(3) Joint method *(Certain ≥ 9)*		(3) Joint method *(Uncertain ≤ 8)*	
Attributes	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*
ASC (Current–Shopping)	−0.987	−7.920	0.687	4.710	−1.033	−22.170	−1.498	−7.570	−0.789	−5.210
ASC (Current–Social)	−0.753	−5.700	0.511	2.590	−0.753	−15.730	−0.612	−3.810	−0.274	−2.720
ASC (Current–Work)	−0.370	−2.840	0.373	1.970	−0.293	−6.340	−0.475	−2.540	−0.376	−7.790
Distance	0.021	17.550	0.063	16.720	0.022	48.550	0.020	12.340	0.017	15.340
Time of day (Night)	3.459	5.260	32.659	7.250	3.540	15.360	2.770	3.170	3.728	4.180
Speeding	−2.608	−2.550	−15.120	−8.330	−2.701	−7.170	−3.593	−2.430	−1.839	−1.770
Travel time	−0.049	−2.660	−0.778	−10.810	−0.042	−6.350	−0.081	−3.160	−0.001	−0.050
Charge	−0.057	−11.460	−0.016	−6.010	−0.058	−30.990	−0.089	−11.080	−0.034	−9.550
Standard deviation	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*
Distance	0.021	17.550	0.063	16.720	0.022	48.550	0.020	12.340	0.017	15.340
Time of day (Night)	3.459	5.260	32.659	7.250	3.540	15.360	2.770	3.170	3.728	4.180
Speeding	2.608	2.550	15.120	8.330	2.701	7.170	3.593	2.430	1.839	1.770
Travel time	0.049	2.660	0.778	10.810	0.042	6.350	0.081	3.160	0.001	0.050
Charge	0.057	11.460	0.016	6.010	0.058	30.990	0.089	11.080	0.034	9.550
GMX	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*		*t-ratio*	
Tau (scale heterogeneity parameter)	0.990	22.840	0.807	27.080	0.995	61.160	0.686		20.690	
Gamma (weighting parameter)	0.500	5.450	1.000	8.590	0.50	15.160	0.500		5.850	
Sigma (mean parameter of scale variance)	0.991		0.994		0.991		0.995			
Model fit										
Sample	456		456		456		456			
Observations	1824		1824		1824		1824			
Log-Likelihood (Base)	−4007.738		−4007.738		−29,750.421		−5272.038			
Log-Likelihood (Model)	−1657.810		−953.696		−12,201.724		−2801.835			
Akaike Information Criterion	1.829		1.057		13.390		3.091			
McFadden ρ^2	0.586		0.762		0.590		0.469			

Note: The constrained the triangular distribution with standard deviation spread set equal to the mean provided the best model fit. T-test for Tau and Gamma are against zero.

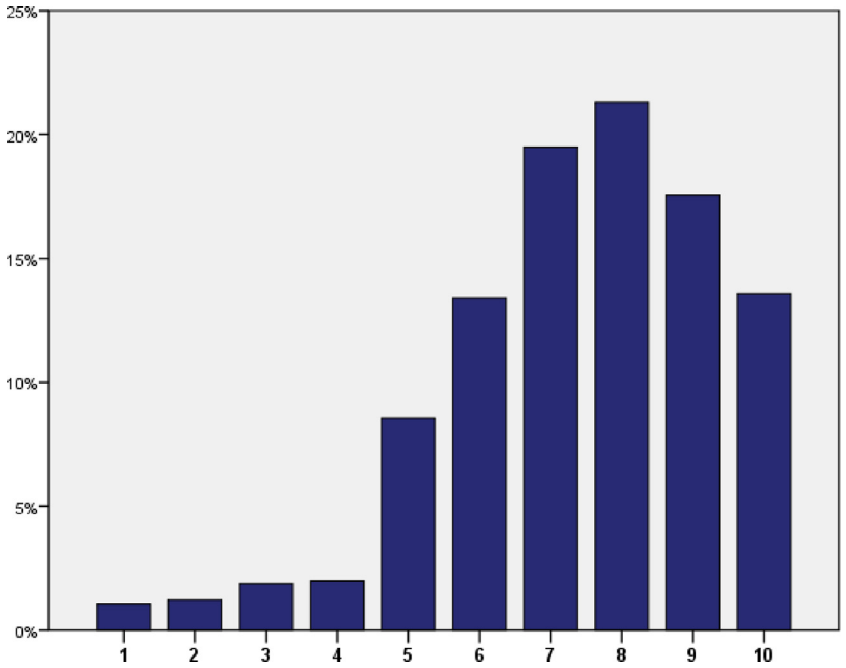

Fig. 3. Certainty Scale Distribution.

utility function. We trialed alternative specific parameters for the attributes themselves, but doing so provided no significant difference in model fit. This was intuitively pleasing because our a priori expectation was that respondents should display a similar sensitivity to the driving behavior attributes irrespective of what type of trip they are making.[4]

Results for distance parameters suggest that participants were concerned with the ability to drive and were reluctant to significantly reduce their driving. Nevertheless, the significant parameter for charge indicates that on average participants preferred to choose trip options with lower charges and were willing to change some of their current driving behavior to reduce the charges in order to make some money. As might be anticipated, participants prefer to maintain their night driving and are travel-time sensitive (i.e., they dislike extra travel time per trip). Interestingly, participants also preferred driving options with less speeding irrespective of any time penalties incurred.

The negative alternative specific constant for the status quo alternative for each of the three trip purposes across three of the four models (base, weighted and joint) suggests that participants were less likely to choose the current alternative and favored the hypothetical alternatives, ceteris paribus. In the model estimated using the recode method, however, the alternative specific constants are significant and positive, indicating that under this method participants are estimated to be more likely to choose the current alternative (in the original data just 27% of respondents selected the status quo, whereas after uncertain responses are recoded back to current behavior 63% of respondents are assumed to select the status quo alternative). In other words, the current alternative is preferred (ceteris paribus) because uncertain responses were recoded into the status quo (i.e. they are uncertain so it is assumed they would not change behavior).

With respect to the joint method, differences between the parameter estimates highlight that difference in impacts arise as a result of whether respondents are certain about their stated choice or not. In particular, instances where the choice is more certain, there is a stronger preference for an alternative proposed driving behavior (that reduces risk) than their current behavior. Also, when respondents are more certain about their stated choice, the proposed reductions in speeding and travel times as well as the impact of the charge applied for exposure to risk all have much bigger impacts on choice than in the case where respondents were uncertain about the alternative they selected in the choice task.

Although parameters across models cannot be directly compared, the parameters for the probability weighted method are very similar in magnitude to the base model parameters. Conversely, the parameters for the recode method are markedly different to base model parameters. In particular, night driving, speeding and travel time parameters are comparatively large in this model. These differences are further highlighted in the resulting hypothetical bias classifications presented in the following section of this paper. Aside from the distance parameter, there are some apparent differences in the remaining parameters between the certain and uncertain alternatives in the joint estimation method model. Assessing the suitability

[4] The test was conducted on the base model using the method of Swait and Louviere (1993), which is fundamentally a Wald test that compare the model fit of the model containing all choice observations with each model estimated on data from each choice scenario (work, shopping, social). The observed value of this test of 17.80 is less than the critical chi-squared value of 18.31 at five per cent level of significance with ten degrees of freedom. This indicates that there is non-significant statistical variation between pooling the data and treating the data separately.

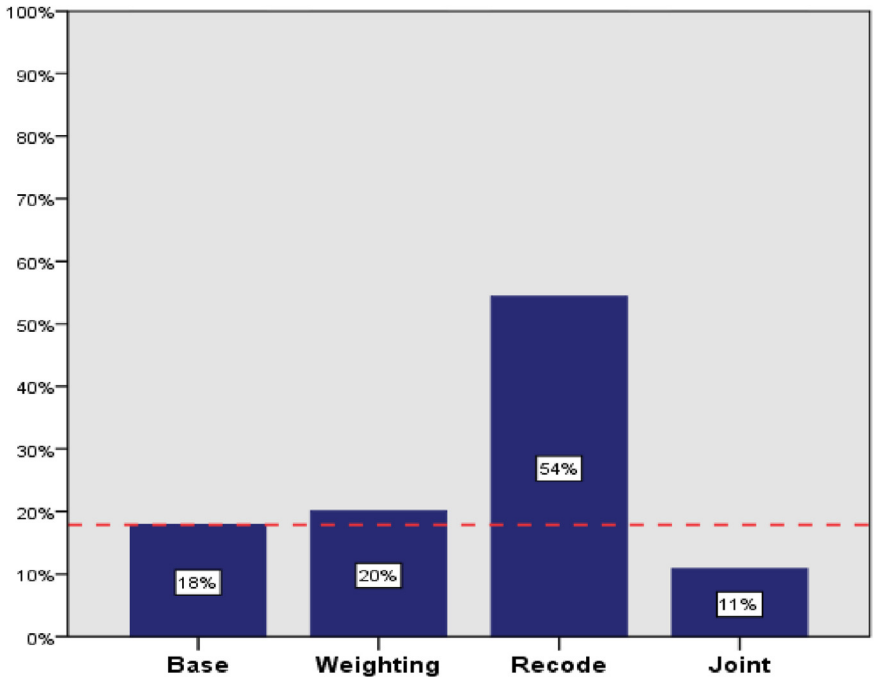

Fig. 4. Hypothetical Bias in Model Prediction.

of the GMX models, the gamma (weighting parameter) and tau (scale heterogeneity parameter) parameters are both significant. The significance of these parameters indicates that the GMX model (via a more flexible distribution) provides an improvement to the Mixed Multinomial Logit (MMNL) model.

4.2. Assessing hypothetical bias mitigation

4.2.1. Model predictions

The total percentage of participants affected by hypothetical bias using the different certainty calibration methods is displayed graphically in Fig. 4. In evaluating the effectiveness of the techniques, each method is compared to the base model (i.e., the percentage of participants affected by hypothetical bias without the use of certainty calibration techniques). Overall, hypothetical bias is an issue for a substantial number of participants in this study, with 18% of respondents affected.

After applying the probability weighting method, 20% of respondents were classified as biased, making little difference from the simple base model estimation. The recode method performs the worst of the three methods tested in this study. In all scenarios the percentage of hypothetical bias is increased dramatically using this method, with more than half the sample being biased when using this approach. This is thought to have arisen because participants who were uncertain about changes in the stated choice survey did not necessarily revert back to their current behavior in the "GPS After" phase; but instead made smaller changes to their driving. The joint estimation method performs the best of the three calibration methods by significantly reducing the extent of hypothetical bias.

A breakdown of how the calibration process affects results is shown in Fig. 5. These graphs isolate just those people who modified their behavior in the "GPS After" phase in response to the exposure-based insurance scheme. In the base scenario where no calibration was applied, eight percent of respondents who actually positively changed driving behavior to reduce risk had stated choice results that indicated they had less than a 25% chance of doing so. Twelve percent of respondents who actually changed where predicted to have a 24–50% chance of doing so. Of those respondents whose stated choice results indicated they had a higher probability of modifying behavior than not (i.e. their hypothetical results did not exhibit bias), 39% had a predicted probability of change between 51–75% and 41% of respondents who actually changed had stated choice results that results in more than a 75% predicted chance of changing in this way.

There is very little change in how results are distributed using the probability weighting method. Dramatically, the recode method results in a significant distributional change, with 45% of respondents who actually changed their driving behavior now predicted to have less than a 25% of actually doing so. Using the method of jointly estimating stated choice and choice certainty means that more than half of respondents who actually changed have a predicted probability of doing so in excess of 75%. Overall it can be seen that not only does the joint method reduce the number of biased participants (lower end of the distribution) but it also increases the overall model predictions (upper end of the distribution).

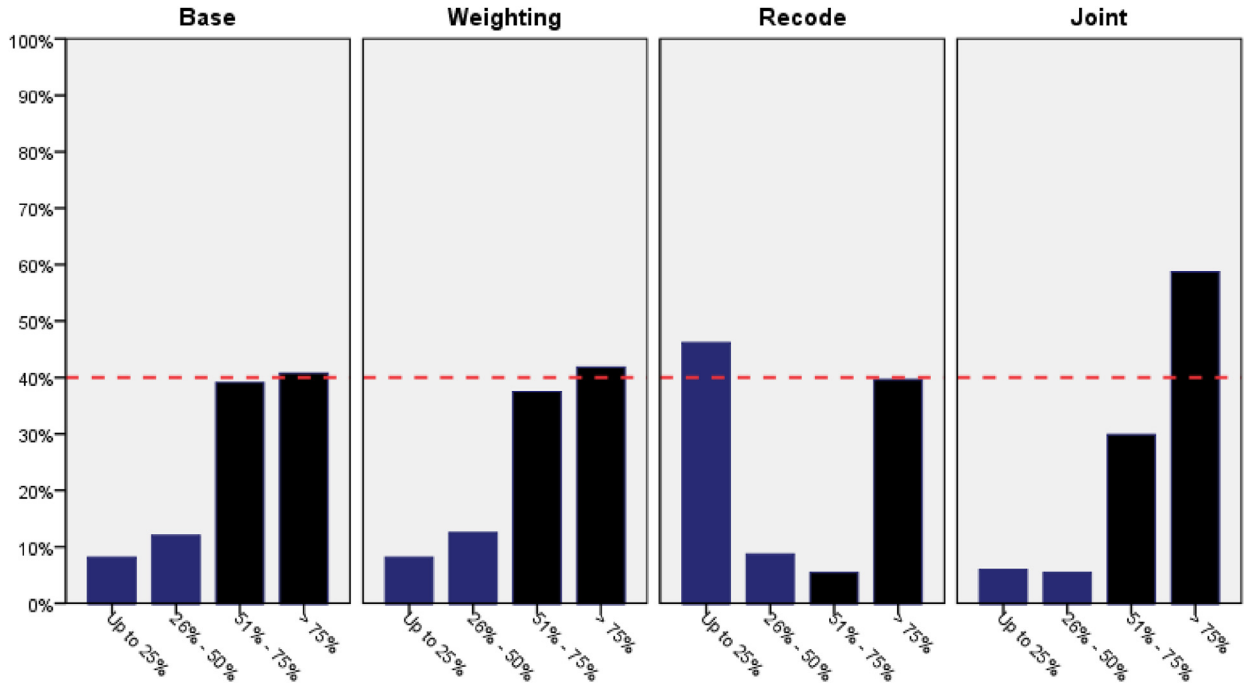

Fig. 5. Distribution of Hypothetical Bias in Model Predictions.

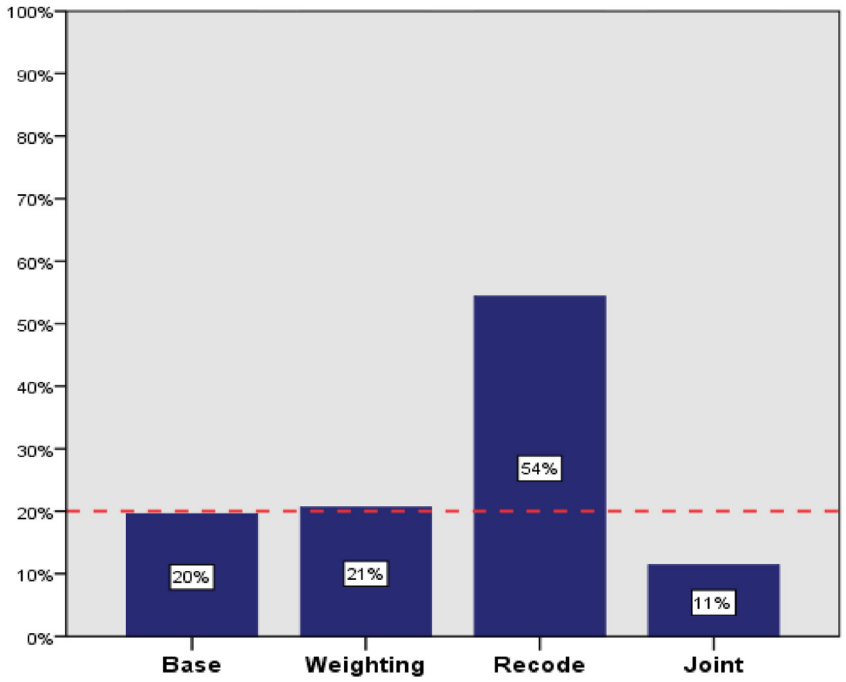

Fig. 6. Hypothetical Bias in Total Willingness to Pay.

4.2.2. Total willingness to pay

The bias classifications generated using TWTP measures are very similar to the model prediction bias classifications (see Fig. 6). Once again the probability weighting method does not change the level of hypothetical bias and the recode method significantly increases hypothetical bias when compared to the base model. Similarly the joint method significantly reduces the prevalence of hypothetical bias, almost having the number of respondents exhibiting this bias. A breakdown of the TWTP distribution for the base model and the three certainty calibration methods is shown in Fig. 7.

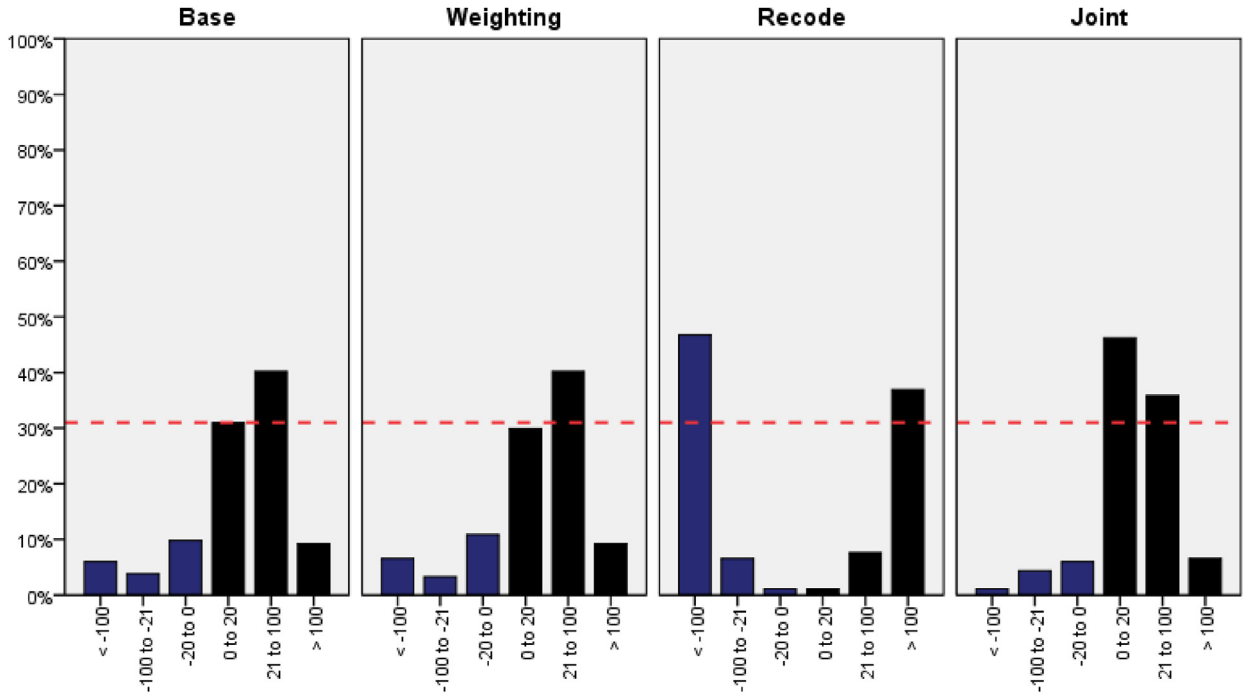

Fig. 7. Distribution of Hypothetical Bias in Total Willingness to Pay.

5. Discussion and conclusion

Despite the importance of hypothetical bias, there are currently only a limited number of research papers which explore this issue in choice experiments. The general consensus from these studies is that choice experiments may, just as with contingent valuation methods, be prone to the phenomenon (Broadbent 2014). One reason for the relative paucity in studies that examine hypothetical bias is that very few studies actually have revealed preference data to compare with their stated choice responses. This is probably a result of two factors: the lack of awareness or acknowledgement of hypothetical bias; and the difficulty, cost and in many cases the inability to collect appropriate revealed data for comparison.

Based on the work by Li and Mattsson (1995), the use of certainty measures as a tool to mitigate hypothetical bias has gained attention in the contingent valuation literature and is, as a consequence, gaining small exposure within the choice literature. Certainty calibration incorporates respondent certainty (uncertainty) with decision-making into the evaluation criteria. It is posited that respondents who are uncertain about their decisions are less likely to follow through with their choices in actuality. At this point in time the number of published instances where certainty calibration has been explored in the context of choice experiments is limited to Champ et al. (1997), Norwood (2005), Brouwer et al. (2010), Ready et al. (2010), Beck et al. (2013), Fifer et al. (2014) and Rose et al. (2015). Additionally, the majority of these studies address the valuation of public goods.

The limited literature and lack of a coherent theoretical basis as to how certainty calibration should be employed led Beck et al. (2013) to conclude that the value of such a technique remains unclear. In response to this, the present study seeks to compare multiple methods of certainty calibration, in particular examining three methods that have recently appeared in the literature: probability weighting of choice sets via the certainty score assigned to each choice; the traditional methodology of recoding uncertain responses to the status quo; and jointly modeling choice and uncertainty. The results from this study support previous findings that participant (un)certainty is clearly associated with hypothetical bias. Alarmingly, in the context of our study of a motor vehicle insurance product we find that one of the more popular methods of certainty scaling, recoding uncertain responses, induces a greater degree of bias than doing nothing at all. This result strengthens the call for further research into certainty scales before they become widely adopted, as the impact of this methodology is itself still uncertain. Our findings also suggest, more positively, that accounting for certainty indexing using a joint estimation approach significantly reduces the extent of hypothetical bias while at the same time accounting for potential endogeneity bias. One limitation to this approach is the extra parameters required for estimation however for most choice experiment applications the inclusion of the extra parameters required using this method will not be prohibitive.

The remaining certainty calibration method tested in this study, probability weighting choices according to choice certainty, was unsuccessful in mitigating hypothetical bias. Conclusions drawn about the efficacy of certainty calibration methods used in this research are limited to the specific methods tested. Future research could address this by comparing a

10-point certainty scale and a categorical or qualitative certainty scale, examining the role of opt-out alternatives or even different messages used to describe how certain respondents are in their choices.

Overall, the practical implications of these findings provide a cautionary warning to researchers and practitioners in this field who use the results from choice models to aid in making important market decisions. We have found that, in the context of a truly hypothetical product where preferences are established in a multi-alternative and multi-attribute choice experiment, hypothetical bias is a significant issue. Our results lead us to urge that more research be done on the existence and sources of hypothetical bias. In this case we made use of a study where a real market could be constructed in order to test the accuracy of the stated choices. Our case study of exposure-based insurance changing is a true hypothetical offering with which respondents have no experience. Other types of markets should also be considered, in particular choices where the respondents do have some familiarity with the choice alternatives. In studying hypothetical bias we are confident that innovative methods to help mitigate it can be found. Positively, we do find several methods of incorporating respondent reported certainty in the estimation process are successful, in particular our results show that when correctly administered the simultaneous estimation of choice and choice certainty has the potential to considerably reduce the prevalence of hypothetical bias. Based on our analysis we recommend that further study be done on alternative methods for simultaneous estimation (for example Dekker et al. 2014) to see if further reductions may be made.

Overall, our findings lead us to recommend that researchers and practitioners include a certainty scale in future studies which employ choice experiments and to test for these effects. While studies of methods to bring stated preference data into alignment with real market behavior are important to choice analysis, perhaps it is time to also revisit the fundamental premise of external validity, of which hypothetical bias is part.

Acknowledgment

The authors would like to acknowledge the comments by two anonymous referees, whose input has improved this paper.

References

Aadland, D., Caplan, A.J., 2003. Willingness to pay for curbside recycling with dection and mitigation of hypothetical bias. American Journal of Agricultural Economics 85, 492–502.
Aadland, D., Caplan, A.J., 2006. Cheap talk reconsidered: new evidence from CVM. Journal of Economic Behavior and Organization 60, 562–578.
Alfnes, F., Guttormsen, A., Steine, G., Kolstad, K., 2006. Consumers' willingness to pay for the color of salmon: a choice experiment with real economic incentives. American Journal of Agricultural Economics 88 (4), 1050–1061.
Arentze, T., Borgers, A., Timmermans, H., Del Mistro, R., 2003. Transport stated choice responses: effects of task complexity, presentation format and literacy. Transportation Research Part E 39 (3), 229–244.
Beck, M.J., Rose, J.M., Hensher, D.A., 2013. Consistently inconsistent: the role of certainty, acceptability and scale in automobile choice. Transportation Research Part E 56, 81–93.
Ben-Akiva, M., Bradley, M., Morikawa, T., Benjamin, J., Novak, T., Oppewal, H., Rao, V., 1994. Combining revealed and stated preferences data. Marketing Letters 5 (4), 335–350.
Blumenschein, K., Blomquist, G.C., Johannesson, M., Horn, N., Freeman, P., 2008. Eliciting willingness to pay without bias: evidence from a field experiment. Economic Journal 118 (525), 114–137.
Bradley, M., Daly, A., 1994. Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. Transportation 21 (2), 167–184.
Brazell J.D. and Louviere J.J. (1998). "Length effects in conjoint choice experiments and surveys: an explanation based on cumulative cognitive burden." Working Paper, University of Sydney, July.
Brouwer, R., Dekker, T., Rolfe, J., Windle, J., 2010. Choice certainty and consistency in repeated choice experiments. Environmental and Resource Economics 46 (1), 93–109.
Brownstone, D., Small, K., 2005. Valuing time and reliability: assessing the evidence from road pricing demonstrations. Transportation Research Part A 39 (3), 279–293.
Carlsson, F., Daruvala, D., Jaldell, H., 2010. Do You do what you say or do you do what you say others do? Journal of Choice Modelling 3 (2), 113–133.
Carson, R.T., Groves Jr., T.F., 2007. Incentive and informational properties of preference questions. Environmental and Resource Economics 37 (1), 181–210.
Caussade, S., Ortuzar, J.D.D., Rizzi, L.I., Hensher, D.A., 2005. Assessing the influence of design dimensions on stated choice experiment estimates. Transportation Research Part B 39 (7), 621–640.
Champ, P.A., Bishop, R.C., 2001. Donation payment mechanisms and contingent valuation: an empirical study of hypothetical bias. Environmental and Resource Economics 19 (4), 383–402.
Champ, P.A., Moore, R., Bishop, R.C., 2009. A comparison of approaches to mitigate hypothetical bias. Agricultural and Resource Economics Review 38 (2), 166–180.
Champ, P.A., Bishop, R.C., Brown, T.C., McCollum, D.W., 1997. Using donation mechanisms to value non-use benefits from public goods. Journal of Environmental Economics and Management 33 (2), 151–162.
Chang, J.B., Lusk, J.L., Norwood, F.B., 2009. "How closely do hypothetical surveys and laboratory experiments predict field behavior?". American Journal of Agricultural Economics 91, 518–534.
Chowdhury, S., Meenakshi, J.V., Tomlins, K.I., Owori, C., 2010. Are consumers willing to pay more for bio-fortified Foods? Evidence from a field experiment in Uganda. American Journal of Agricultural Economics 93 (1), 83–97.
Cummings, R.G., Brookshire, S., Schulze, W.D., 1986. Valuing Environmental Goods: A State of the Arts Assessment of the Contingent Method. Rowman and Allanheld, Totowa, NJ.
Cummings, R.G., Taylor, L.O., 1999. Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method. American Economic Review 89 (3), 649–666.
Davies, S.P., Loomis, J.B., 2010. An improved method for calibrating purchase intentions in stated preference demand models. Journal of Agricultural and Applied Economics 42, 679–693.
Dekker, T., Hess, S., Brouwer, R., Hofkes, M., 2014. Implicitly or Explicitly Uncertain?" Working Paper. University of Leeds, United Kingdom.
DeShazo, J.R., Fermo, G., 2002. Designing choice sets for stated preference methods: the effects of complexity on choice consistency. Journal of Environmental Economics and Management 44 (1), 123–143.
Diamond, P.A., Hausman, J.A., 1994. Contingent valuation: is some number better than no number? Journal of Economic Perspectives 8 (4), 45–64.
Ding, M., Grewal, R., Liechty, J., 2005. Incentive-aligned conjoint analysis. Journal of Marketing Research 42 (1), 67–82.

Ethier, R.G., Poe, G.L., Schulze, W.D., Clark, J., 2000. A comparison of hypothetical phone and mail contingent valuation responses for green-pricing electricity programs. Land Economics 76 (1), 54–67.
Fiebig, D.G., Keane, M.P., Louviere, J., Wasi, N., 2010. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. Marketing Science 29 (3), 393–421.
Fifer, S., Rose, J.R., Greaves, S.P., 2014. Hypothetical bias in stated choice experiments: is it a problem? and if so, how do we deal with it? Transportation Research Part A 61, 164–177.
Fifer, S., Greaves, S.P., Rose, J.M., Elisson, R., 2011. A combined GPS/Stated choice experiment to estimate values of crash-risk reduction. Journal of Choice Modelling 4 (1), 44–61.
Fox, J.A., Shogren, J.F., Hayes, D.J., Kliebenstein, J.B., 1998. Calibrating contingent values with experimental auction markets. American Journal of Agricultural Economics 80 (3), 455–465.
Greaves, S.P., Fifer, S., 2010. Development of a kilometer-based rewards system to encourage safer driving practices. Transportation Research Record: Journal of the Transportation Research Board 2182, 88–96.
Greene, W.H., Hensher, D.A., 2010. Does Scale heterogeneity across individuals matter? An empirical assessment of alternative logit models. Transportation 37 (3), 413–428.
Haab, T.C., McConnell, K.E., 2002. Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation. Edward Elgar, Northampton, United Kingdom.
Hanemann, W.M., 1994. Valuing the environment through contingent valuation. Journal of Economic Perspectives 8 (4), 19–43.
Harrison, G.W., 2006. Experiment evidence on alternative environmental valuation methods. Environmental and Resource Economics 34, 125–162.
Harrison, G.W., 2007. Making choice studies incentive compatible. In: Kanninen, B. (Ed.), Valuing Environmental Amenities Using Stated Choice Studies. Springer, The Netherlands, pp. 67–110.
Harrison, G.W., Harstad, R.M., Rutstrom, E.E., 2004. Experimental methods and elicitation of values. Experimental Economics 7, 125–162.
Harrison, G.W., Rutstrom, E.E., 2008. Experimental evidence on the existence of hypothetical bias in value elicitation methods. Handbook of Experimental Economics 1, 752–767.
Hensher, D.A., 2010. Hypothetical bias, choice experiments and willingness to pay. Transportation Research Part B 44 (6), 735–752.
Hensher, D.A., Bradley, M., 1993. Using stated response data to enrich revealed preference discrete choice models. Marketing Letters 4 (2), 139–152.
Hensher, D.A., Rose, J.M., Greene, W.H., 2008. Combining RP and SP data: biases in using the neste logit 'trick' – constrasts with flexible mixed logit incorporating panel and scale effects. Journal of Transport Geography 16 (2), 126–133.
Herriges, J.A., Kling, C.L., Azevedo, C.D., 1999. Linking revealed and stated preferences to test external validity. Working Paper 222 Iowa State University; Center for Agricultural and Rural Development http://www.card.iastate.edu/publications/DBS/PDFFiles/99wp222.pdf, 15/12/15.
Herriges, J.A., Kling, C.L., Liu, C., Tobias, J., 2010. What are the consequences of consequentiality? Journal of Environmental Economics and Management 59 (1), 67–81.
Hess, S., Rose, J.M., 2012. Can scale and coefficient heterogeneity be separated in random coefficients models? Transportation 39 (6), 1225–1239.
Hudson, D., Gallardo, K., Hanson, T., 2012. A comparison of choice experiments and actual grocery store behavior: an empirical application to seafood products. Journal of Agricultural and Applied Economics 44 (1), 49–62.
Isacsson G. (2007). "The Trade Off Between Time and Money: Is There a Difference Between Real and Hypothetical Choices?" Swedish National Road and Transport Research Institute, Borlange, Sweden.
Jacquemet, N., Joule, R., Luchini, S., Shogren, J.F., 2013. Preference elicitation under oath. Journal of Environmental Economics and Management 65, 110–132.
Johannesson, M., Blomquist, G.C., Blumenschein, K., Johansson, P., Liljas, B., Conor, R.M.O'., 1999. Calibrating hypothetical willingness to pay responses. Journal of Risk and Uncertainty 18 (1), 21–32.
Keane, M., The Generalized Logit Model: Preliminary Ideas on a Research Program, Motorola-CenSoC Hong Kong Meeting, October 22, 2006.
Lancsar, E., Swait, J., 2014. Reconceptualising the external validity of discrete choice experiments. PharmacoEconomics 32 (10), 951–965.
Landry, C.E., List, J.A., 2007. Using ex ante approaches to obtain credible signals for value in contingent markets: evidence from the field. American Journal of Agricultural Economics 89, 420–429.
Li, C., Mattsson, L., 1995. Discrete choice under preference uncertainty: an improved structural model for contingent valuation. Journal of Environmental Economics and Management 28 (2), 256–269.
List, J.A., 2001. Do explicit warnings eliminate the hypothetical bias in elicitation procedures? evidence from field auctions for sportscards. American Economic Review 91, 1498–1507.
List, J.A., Gallet, C., 2001. What experimental protocol influence disparities between actual and hypothetical stated values? Environmental and Resource Economics 20 (3), 241–254.
List, J.A., Sinha, P., Taylor, M.H., 2006. Using choice experiments to value non-market goods and services: evidence from field experiments. Advances in Economic Analysis and Policy 6 (2) 1132-1132.
Little, J., Berrens, R.P., 2004. Explaining disparities between actual and hypothetical stated values: further investigation using meta-analysis. Economics Bulletin 3 (6), 1–13.
Loomis, J.B., 2011. What's to know about hypothetical bias in stated preference valuation studies? Journal of Economic Surveys 25 (2), 363–370.
Loomis, J.B., 2014. Strategies for overcoming hypothetical bias in stated preference surveys. Journal of Agricultural and Resource Economics 39 (1), 34–46.
Loomis, J., Pierce, C., Manfredo, M., 2000. Using the demand for hunting licenses to evaluate contingent valuation estimates of willingness to pay. Applied Economics Letters 7, 435–438.
Louviere, J.J., 1988. Conjoint analysis modelling of stated preferences: a review of theory, methods, recent developments and external validity. Journal of Transport Economics and Policy 22 (1), 93–119.
Lucas, J.W., 2003. Theory-testing, generalization, and the problem of external validity. Sociological Theory 21 (3), 236–253.
Lusk, J.L., Schroeder, T.C., 2004. Are choice experiments incentive compatible? a test with quality differentiated beef steaks. American Journal of Agricultural Economics 86 (2), 467–482.
Miller, K.M., Hofstetter, R., Krohmer, H., Zhang, Z.J., 2011. How should we measure consumers' willingness to pay? an empirical comparison of state-of-the-art approaches. Journal of Marketing Research 48 (1), 172–184.
Mitchell, R.C., Carson, R.T., 1989. Using Surveys to Value Public Goods: The Contingent Valuation Method. John Hopkins University Press, Baltimore, MD, United States.
Moore, R., Bishop, R.C., Provencher, B., Champ, P.A., 2010. Accounting for respondent uncertainty to improve willingness-to-pay estimates. Canadian Journal of Agricultural Economics / Revue Canadienne d'Agroeconomie 58 (3), 381–401.
Morrison, M., Brown, T.C., 2009. Testing the effectiveness of certainty scales, cheap talk and dissonance-minimization in reducing hypothetical bias in contingent valuation studies. Environmental and Resource Economics 44 (3), 307–326.
Moser, R., Notaro, S., Raffaelli, R., 2010. Using your own money makes the difference: testing the hypothetical bias with a real choice experiment. In: Proceedings of 2010 World Congress of Environmental and Resource Economics. Montreal, Canada.
Murphy, J.J., Allen, P.G., Stevens, T.H., Weatherhead, D., 2005. A meta-analysis of hypothetical bias in stated preference valuation. Environmental & Resource Economics 30 (3), 313–325.
Myerson, R.B., 1979. Incentive compatibility and the bargaining problem. Econometrica 47 (1), 61–73.
Norwood, F.B., 2005. Can calibration reconcile stated and observed preferences? Journal of Agricultural and Applied Economics 37 (1), 237–248.
Olsson B. (2005). "Accounting for Response Uncertainty in Stated Preference Methods." Paper presented at the EAERE Congress, Bremen, Germany.
Manski, C.F., Lerman, S.R., 1977. The estimation of choice probabilities from choice based samples. Econometrica 45 (8), 1977–1988.
Portney, P.R., 1994. The contingent valuation debate: why economists should care. Journal of Economic Perspectives 8 (4), 3–17.

Ready, R.C., Champ, P.A., Lawton, J.L., 2010. Using respondent uncertainty to mitigate hypothetical bias in a stated choice experiment. Land Economics 86 (2), 363–381.
Rose, J.M., Beck, M.J., Hensher, D.A., 2015. The joint estimation of respondent-reported certainty and acceptability with choice. Transportation Research Part A 71, 141–152.
Rose, J.M., Bliemer, M.C.J., 2009. Constructing efficient stated choice experimental designs. Transport Reviews 29 (5), 587–617.
Rose, J.M., Bliemer, M.C.J., Hensher, D.A., Collins, A.T., 2008. Designing efficient stated choice experiments in the presence of reference alternatives. Transportation Research Part B 42, 395–406.
Samuelson, P.A., 1955. Diagrammatic exposition theory of public expenditure. Review of Economics and Statistics 37, 350–356.
Stevens, T.H., Tabatabaei, M., Lass, D., 2013. Oaths and hypothetical bias. Journal of Environmental Management 127, 135–141.
Swait, J., Adamowicz, W., 2001. The influence of task complexity on consumer choice: a latent class model of decision strategy switching. Journal of Consumer Research 28 (1), 135–148.
Swait, J., Louviere, J.J., 1993. "The role of the scale parameter in the estimation and comparison of multinomial logit models. Journal of Marketing Research 30, 305–314.
Train, K., 2009. Discrete Choice Methods with Simulation, 2nd Edition Cambridge University Press, Cambridge, United Kingdom.
Vossler, C.A., Doyon, M., Rondeau, D., 2012. Truth in consequentiality: theory and field evidence on discrete choice experiments. American Economic Journal: Microeconomics 4, 145–171.
Vossler, C.A., Evans, M.F., 2009. Bridging the gap between the field and the lab: environmental goods, policy maker input, and consequentiality. Journal of Environmental Economics and Management 58 (3), 338–345.
Wardman, M., Whelan, G., 2001. Valuation of improved railway rolling stock: a review of the literature and new evidence. Transport Reviews 21 (4), 415–448.
Wardman, M., Shires, J., 2001. Comparison of within mode revealed preference and stated preference choice models. In: Paper presented at European Transport Conference. Cambridge, United Kingdom.
Whitehead, J.C., Cherry, T.L., 2007. Willingness to pay for a green energy program: a comparison of ex-ante and ex-post hypothetical bias mitigation approaches. Resource and Energy Economics 29, 247–261.
Wilson, R., 1978. Information, efficiency, and the core of an economy. Econometrica 46 (4), 807–816.