I T L S

WORKING PAPER

ITLS-WP-09-01

Hypothetical bias, choice experiments and willingness to pay

By

David A Hensher

January 2009

ISSN 1832-570X

INSTITUTE of TRANSPORT and LOGISTICS STUDIES

The Australian Key Centre in Transport and Logistics Management

The University of Sydney

Established under the Australian Research Council's Key Centre Program.

NUMBER: Working Paper ITLS-WP-09-01

TITLE: **Hypothetical bias, choice experiments and willingness to pay**

ABSTRACT:

There is growing interest in establishing the extent of differences in willingness to pay (WTP) for attributes, such as travel time savings, that are derived from real market settings and hypothetical (to varying degrees) settings. Non-experiment external validity tests involving observation of choice activity in a natural environment, where the individuals do not know they are in an experiment, are rare. In contrast the majority of tests are a test of external validity between hypothetical and actual experiments. Deviation from real market evidence is referred to in the literature broadly as hypothetical bias. The challenge is to identify such bias, and to the extent to which it exists, establishing possible ways to minimise it. This paper reviews the efforts to date to identify and 'calibrate' WTP derived from one or more methods that involve assessment of hypothetical settings, be they (i) contingent valuation methods, (ii) choice experiments involving trading attributes between multiple alternatives, with or without referencing, or (iii) methods involving salient or non-salient incentives linked to actual behaviour. Despite progress in identifying possible contributions to differences in marginal WTP, there is no solid evidence, although plenty of speculation, to explain the differences between all manner of hypothetical experiments and non-experimental evidence. The absence of non-experimental evidence from natural field experiments remains a major barrier to confirmation of under or over-estimation. We find, however, that the role of referencing of an experiment relative to a real experience (including evidence from revealed preference (RP) studies), in the design of choice experiments, appears to offer promise in the derivation of estimates of WTP that have a meaningful link to real market activity, closing the gap between RP and SC WTP outputs.

KEY WORDS: ***Hypothetical bias, willingness to pay, choice experiments, contingent valuation, contextual bias, referencing, revealed behaviour***

AUTHORS: David A Hensher

CONTACT:

Institute of Transport and Logistics Studies **(C37)**

The Australian Key Centre in Transport Management

The University of Sydney NSW 2006 Australia

Telephone:	+61 9351 0071
Facsimile:	+61 9351 0088
E-mail:	itlsinfo@itls.usyd.edu.au
Internet:	http://www.itls.usyd.edu.au

DATE: January 2009

Acknowledgements.
Discussions with Glenn Harrison, Ken Small, David Brownstone, John Rose, Ken Train, Mike Bliemer, David Layton and Sean Puckett have been invaluable in the preparation of this paper. I especially thank Glenn Harrison for clarifying a number of points in earlier versions about contingent value studies. The detailed comments of three referees are also appreciated, even though some points remain contentious for one referee. This paper has benefited from presentations in The Netherlands as a plenary paper at the European Transport Conference on 7 October 2008 and a seminar at Significance and NEA. This research is funded by the Australian Research Council Discovery Project Grant DP0770618.

1. Introduction

The extent to which individuals might behave inconsistently, when they do not have to back up their choices with real commitments, is linked to the notion of hypothetical bias, and is becoming a major question in transportation research as we build up a substantial portfolio of empirical evidence on estimates of willingness to pay (WTP) for specific attributes from choice experiments (CE's).[1]

In the context of travel behaviour, we have one influential paper by Brownstone and Small (2005) which concludes with the suggestion, based on a toll road context in California, that there are significant differences between revealed preference (RP) and stated choice marginal willingness to pay when the latter is obtained from choice experiments. RP studies can vary from those based on actual market evidence or revealed behaviour (e.g., Brownstone and Small 2005), to RP experiments (e.g., Isacsson 2007), through to traditional discrete choice studies based on a known choice and one or more non-chosen alternatives defined on a respondent's perceptions of attribute levels; or some synthesis rule based on physical networks (as in transport studies), or averaging of attribute levels of non-chosen alternatives using information from individuals who chose a particular alternative (e.g., with a common origin and destination in a trip making context).

Efforts to study the influence of hypothetical bias on marginal WTP ((MWTP)[2] and total WTP (TWTP)[3] in a choice experiment context have been confined largely, but not exclusively, to agricultural and resource applications (see Alfnes and Steine 2005, Alfnes et al. 2006, Lusk and Schroeder 2004, and Carlsson and Martinsson 2001). The evidence is mixed in respect of the relative marginal WTP as one external validity test of a choice model's behavioural realism. For example, Carlsson and Martinsson (2001) and Lusk and Schroeder (2004), in comparing preferences between a hypothetical and actual choice experiment, found no evidence of difference in the MWTP; in contrast Isacsson (2007), in the context of trading time with money, found that the MWTP based on the hypothetical experiment was almost 50 percent lower at the mean than the real experiment MWTP, supporting the conclusions by Brownstone and Small (2005) in a transport context that "…the value of time saved on the morning commute is quite high (between $20 and $40 per hour) when based on revealed behavior, and less than half that amount when based on hypothetical behavior" (page 279).

Lusk and Schroeder (2004) and Alfnes and Steine (2005) found significant differences in TWTP[4], calculated by comparing a null alternative with an application scenario. Carlsson and Martinsson (2001) did not investigate TWTP because they did not have an opt-out or 'none of these' alternative (described by some researchers as a serious design

[1] We use the phrase *choice experiment* to refer to the methods commonly adopted in transportation studies to evaluate packages of attributes, referred to as alternatives, and to then make a choice or to rank order the alternatives.

[2] We use the phrase 'marginal willingness to pay' to refer to the valuation of a specific attribute.

[3] Total WTP is a language common in health, environmental and resource studies to represent the change in total consumer surplus between the null alternative and the application of interest. The estimate is based on the total utility difference in dollars of a base alternative and a scenario where an attribute takes a specific value (e.g., unconstrained mouse hunting vs. banning mouse hunting).

[4] The literature in agricultural, resource and environmental valuation does not calibrate the alternative-specific constants, even when the application has real market alternatives with known market share (e.g. Lusk and Schroeder 2004). This may in part explain the significant differences in the total WTP in contrast to the non-significant differences in MWTP.

flaw[5] – see Harrison 2006), and hence they forced respondents to make a choice that included some level of payment. Although not conclusive, the literature 'suggests' that the presence or absence of the opt-out or 'no choice' alternative does often make a noticeable difference to the evidence. For example, Ladenburg et al. (2007), based on a review of the evidence from a number of studies and their own empirical investigations, conclude:

> "Assuming that the observed effect of the opt-out reminder reflects that the stated preferences are brought closer to the true preferences, adding a relatively short-scripted opt-out reminder ……. will effectively reduce hypothetical bias further."

Although this specific source of potential hypothetical bias needs consideration, the evidence from many quality studies is well summarized in Murphy et al. (2005):

> "…it is likely that a number of factors affect hypothetical bias and therefore no single technique will be the magic bullet that eliminates this bias" page 317.

An assessment of the evidence, including meta analyses (e.g., List and Gallet 2001, Murphy et al. 2004) points to a number of potentially key influences on the findings in respect of MWTP and TWTP. These include the nature of the good being studied (private or public), any connotations in terms of environmental consciousness (feel good, yeah-say), the presence or absence of an opt-out alternative, the opportunity to calibrate alternative-specific constants on all or a subset of alternatives that are observed in actual markets, the role of supplementary data to condition the choice outcome, hypothetical or real, that can be encapsulated in the notion of information processing (e.g., identifying heuristics such as imposition of threshold on the way one processes attributes or ignoring certain attributes that impact on choices that may be generic to an individual or specific to the experimental circumstance, including referencing[6] around a known experience[7] - see Hensher 2006, 2008, Rose et al. 2008), and in items that identify 'the confidence with which they would hypothetically purchase the good at the stated alternative or attribute' (Harrison 2007).

The focus of the paper is on the marginal WTP evidence from choice experiments in the transportation context, and the extent to which evidence on hypothetical bias from a wider literature, including the more extensive literature on contingent valuation, can offer guidance on how CE's might be structured to narrow the gap between actual market WTP and WTP derived from hypothetical choice experiments.

This paper is organised as follows. In the next section we present a number of key themes to highlight approaches used to estimate MWTP, the main focus of this paper, and to identify possible sources of hypothetical bias revealed in the literature. This is followed by a limited empirical assessment, using a number of traditional RP and CE data sets, given the absent of non-experiment real choices observed in a natural

[5] The recognition of the role of the opt-out or null alternative has been described by Glenn Harrison as a potentially key insight into why conjoint choice experiments *may* allow analysts to do tight statistical calibration for hypothetical bias (personal communication 9 February 2008).

[6] Referencing is the extent to which an application has an identifiable real observation to benchmark against (e.g., choice amongst existing tolled and free routes used to establish market shares and MWTP for time savings), in contrast to valuation of specific attributes such as noise and safety where a real observation of MWTP is not usually known or able to be assessed unambiguously.

[7] The use of referencing the choice experiment design to a real activity, as in toll road studies, is generally lacking in the literature outside of transportation. Glenn Harrison makes the valid point that this may be a two-edged frame, biasing responses. One way of establishing the presence of bias is to incorporate the reference into the design as a treatment which is present and absent across the within-subject experiment. This is also a way of assessing endogeneity (see Train and Wilson 2008).

environment,[8].designed to suggest directional influences of specific choice experiment (CE) elements on the gap between RP and CE MWTP. We then consider the role of the numerator and denominator in the empirical estimation of MWTP, which suggests that a closer look at referencing within a choice experiment, and how it is handled in model estimation, may be an important way of grounding the choice design in reality, especially where the travel activity is characterised by habitual behaviour, such as commuting, and in helping to establish MWTP that are closer to the evidence from real market choice (i.e., revealed behaviour). The paper concludes with suggestions on where future empirical efforts should be focused if we are to have confidence in empirical evidence from choice experiment studies that can be used as if it were obtained 'at a distance' from observing behaviour in real markets. Contrasts with empirical evidence from traditional revealed preference models can be used as a reference benchmark where it is believed that this is close to real market WTP, enabling us to gauge the extent to which specific treatments of SC WTP might close the gap and reduce hypothetical bias[9].

2. Key themes

This section draws together key assumptions and approaches used in the empirical identification of MWTP. To make some sense of the literature, we begin with a classification based on the nature of the data (real, experimental); followed by a discussion of the key behavioural paradigm (CV, CE); the role of specific features linked in general to uncertainty such as opt-out, cheap-talk and referencing; and efforts to introduce calibration and bias functions to reduce hypothetical bias.

2.1 Data spectrum

The terms revealed, real and actual values are typically used interchangeably in the literature to refer to situations in which an individual makes a consequential economic commitment. In experimental studies, this typically involves payment for a good or service by the participant. Most studies of hypothetical bias assume that these cash-based estimates are unbiased. On the other hand, stated or hypothetical values refer to survey responses that lack any salient economic commitment and typically a hypothetical scenario[10]. A review of the literature suggests that we can reasonably classify evidence on MWTP in terms of three broad categories, each with a set of underlying variations. The classification suggests that the term 'revealed preference' (RP), common in transportation studies, is not strictly equivalent to 'real market data'. The choice response in RP studies may be a real market response, but the data typically imposed on each respondent for non-chosen alternatives is controversial (as is concern in general about lack of variance and measurement error),[11] and it is one reason,

[8] Where the individuals do not know they are in an experiment (see Harrison and List 2004) or what I refer to as 'at a distance',

[9] One referee provided extensive comment in support of conventional RP choice models as evidence of WTP in real markets. This is controversial; however given the extensive use of RP models as if they reproduce real market trading amongst attributes, we have incorporated this benchmark in the revised version of the paper.

[10] One way of distinguishing salient and non-salient circumstances is that a salient economic commitment would be consistent with "I prefer X and Y and I actually chose X". A non-salient economic commitment would be "I prefer X to Y but there is no guarantee I will actually choose X".

[11] As Ben-Akiva et al. (1994) state "the possibilities to elicit real WTP measures are limited because they can be varied only on a small scale".

amongst others, that stated choice methods have blossomed for marginal WTP identification.

Real markets, where we can *observe* actual behaviour, are able to identify the levels of attributes offered by each available alternative, as perceived by the decision making unit. A distinction within the category is based on the presence or absence of an experimental treatment. The *Non-Experimental Focus* entails non-artefactual or anonymous observation of individuals 'at a distance' (i.e., without having to ask any questions; in a sense the individual does not know they are being studied), and recording their choice and attribute levels of relevant alternatives[12]. While this avoids any experimental bias, it often exhibits measurement error associated, in particular, with the non-chosen alternative(s) (e.g., choice between a tolled and free route, between modes of transport), although examples in areas such as choice of quality beef in a supermarket (see Lusk and Schroeder 2004) can avoid measurement error through unambiguous labelling of the attributes of each alternative on offer. This focus, however, is unlikely to be rich enough to observe the processing strategy adopted in respect to attributes and alternatives, which does influence the MWTP (Hensher 2008), including the identification of the relevant choice set from which the observed outcome is obtained. However, as shown in a number of recent papers (e.g. Hensher and Greene 2008, Hensher and Layton 2008), it is possible to infer the probability of a processing rule being invoked without having to ask each individual[13].

The *Experimental Focus* involves giving individuals money or having them earn it to undertake real choices and actions. There is typically a baseline participation fee, together with a sum of money and other attributes that vary according to the alternative chosen (e.g., Isacsson 2007, Lusk and Schroeder 2004). A concern with this focus is that the participation fee may be driving the outcome, as distinct from revealing true behaviour under circumstances where the financial means is internally derived by the individual and is a real trade, in terms of opportunity cost. Carlsson and Martinson 2001 suggest that:

> "…. our test of external validity may not be seen as a test of truthful revelation but rather as a test of external validity between hypothetical and actual experiments." (Carlsson and Martinsson 2001)

Revealed preference (RP) data, common in transportation and location studies, is based on surveying individuals and asking them to describe a recent or current actual alternative, and one or more non-chosen (possibly non-experienced) alternatives with/without any information to identify the attribute processing strategies used. Whereas the actual choice outcome is known, the levels of the attributes are either reported by the respondent or drawn from a synthetic source, such as a modal network in transportation studies. Either way the information associated with the non-chosen

[12] For example, in a high occupancy toll lane (HOT) context in California, the analyst is able to measure the travel times using third party methods (e.g. car following), and identifying the toll as a posted price, as was the situation reported in Brownstone and Small (2005).

[13] It is not yet clear whether the analytical methods implemented to identify the use of various process rules up to a probability are an improvement on the self-stated supplementary questions asked of respondents as to how they processed the attribute data in choice experiments (e.g., non-attended to specific attributes, added up common-metric unit attributes).

alternatives is potentially subject to non-marginal errors in respect of what levels would be experienced if an individual were to choose that alternative in a real market[14].

Stated Choice (SC) data can be classified into two broad classes – contingent valuation (CV)[15] and choice experiments (CE). Both are methods of estimating the non-market value of attributes or amenities such as values of endangered species, recreational or scenic resources, air pollution and travel time savings. These measured values are generally based on the willingness to pay for improved attribute or amenity levels. In the CV context, there are several questions that have been used, both in a controlled experiment and in practice. Among those are dichotomous, open ended, payment card and bidding games. Dichotomous choice is a bid offered to the respondent that he/she can accept or reject, while in the open-ended question the respondent is asked for his or her maximum willingness to pay for something that is of most interest, such as an improved quality environment. The payment card is a mode of question shown to the respondent with several bids printed on it. The respondent is asked if any of those bids is close to their maximum willingness to pay. Finally, the bidding game refers to the sequence of bids offered to the respondent so that his or her maximum willingness to pay can be elicited (Frykblom 1997).

The choice experiment approach (see Rose and Bliemer 2007 for a recent review) is based on surveying individuals using a variety of instruments (e.g., pencil and paper, computer aided personal survey instrument (CAPI), internet-based survey) and asking them to assess a set of analyst-defined alternatives, and to express their preferences through first preference rank (i.e., choosing one), full or partial ranking, and rating all alternatives, with/without analyst knowledge of attribute processing strategy (APS). There are many variants including labelled or unlabelled choice sets (attributes and alternatives) – sometimes referred to as multiple price lists in experimental economics which, when generalised to many attributes, are referred to as multiple attribute lists (equivalent to multiple price lists – see Harrison 2007). These can also have variants such as referenced alternatives based on a pivot around a real action (which are increasingly common in transport studies – see Rose et al. 2008), and allowance for attribute processing, exogenously or endogenously (see Tversky and Kahneman 1981, Hensher 2008, Swait 2001) with various heuristics to define attribute boundaries (Hensher and Layton 2008).

[14] To investigate the possibility of bias caused by systematic misperception of travel times, Ghosh (2001) used perceived time savings to help explain route, mode, and transponder choice in a tolled vs. non-tolled lane choice setting. Perception error (defined as perceived minus actual time savings) was added as an explanatory variable. He found that commuters with larger positive perception errors are more likely to use the toll facility; however, the RP values of time savings are not changed by including this variable, suggesting that RP results may not be affected by perceptual problems. Ghosh was not able to identify whether or not SP results are so affected (see also Brownstone and Small 2005).

[15] The term was apparently first introduced in 1947 by S.V. Ciriacy-Wantrup, who thought that the appropriate procedures employed interviews in which subjects are "asked how much money they are willing to pay for successive additional quantities of a collective extra-market good." One implicit assumption of this definition is that contingent value is not needed for ordinary market goods. But with respect to those goods that are not bought and sold, some device as to replace the set of prices that markets happily make explicit. Toward that end the tester prepares an array of questions about some particular subject matter in order to elicit how much they would be prepared to pay—the so-called WTP—in order to secure the provision of some public good. Alternatively, they are asked how much money someone would have to pay them—the so-called WTA—to discontinue some public project that they hold dear. There is sharp disagreement as to how useful the best of these studies are in making value determinations for widespread studies on the valuation of a full set of public goods—the creation of a national park, the preservation of wildlife in an estuary, the control of epidemics, the pursuit of national security, or whatever.

2.2 Evidence from contingent valuation to guide choice experiments

Although the main focus of this paper is on MWTP and choice experiments, in which the empirical investigation of hypothetical bias is somewhat limited, there is much to be learnt from CV studies in guiding the design and application of CE's and any supplementary data that sets the context and conditions the choice response. There is a burgeoning literature on hypothetical bias associated with contingent valuation (see for example Portney 1994, Hanemann 1994, Diamond and Hausman 1994), summarised in Harrison (2006, 2007) and Carson et al. (1996) under the broad heading of making choice experiments incentive compatible[16]. Much of the focus in CV studies is on assessing hypothetical bias in TWTP; however there are some useful signals on what may be sources of hypothetical bias in estimation of MWTP which can be taken into account in estimation of MWTP in choice experiments.

2.2.1 CV evidence

The method of CV has been the subject of heavy criticism, with much of this debate focusing on the validity of the results, in particular the hypothetical nature of the experiments (see Carson et al. 1996). The accumulating evidence (in the list of references herein) suggests that individuals in such hypothetical CV studies exaggerate their TWTP and MWTP for private and public goods. Several attempts have been made to reduce the influence of this hypothetical bias. Cheap talk scripts seemed to be one of the most successful attempts. Initially suggested by Cummings et al. (1995, 1995a), cheap talk is an attempt to bring down the hypothetical bias by describing and discussing the propensity of respondents to exaggerate stated WTP for a specific good at a specific price. Using private goods, classroom experiments, or closely controlled field settings, cheap-talk proved to be potentially successful (see Cummings and Taylor 1999). While the hypothetical mean TWTP without cheap-talk was significantly higher than TWTP using actual economic commitments, the hypothetical TWTP with cheap-talk script could not be shown to be statistically significantly different from the actual TWTP. In general we would conclude that the evidence is mixed and the debate is still wide open.

List and Gallet (2001)[17] used a meta-analysis to explore whether there are any systematic relationships between various methodological differences and hypothetical bias. Their results indicate that the magnitude of hypothetical bias was statistically less for (a) willingness-to-pay (WTP) as compared to willingness-to-accept (WTA) applications, (b) private as compared to public goods, and (c) one elicitation method, the first price sealed bid, as compared to the Vickrey second price auction baseline.

Murphy et al. (2004) also report the results of a meta-analysis of hypothetical bias in 28 CV studies that report willingness-to-pay and used the same mechanism for eliciting both hypothetical and actual values. The 28 papers generated 83 observations with a median ratio of hypothetical to actual value of 1.35, and a distribution with a severe positive skewness. They found that a choice-based elicitation mechanism, such as

[16] A process is said to be incentive compatible if all of the participants fare best when they truthfully reveal any private information the mechanism asks for. As an illustration, voting systems which create incentives to vote dishonestly *lack* the property of incentive compatibility. In the absence of dummy bidders, a second price auction is an example of mechanism that *is* incentive compatible. There are different degrees of incentive compatibility: in some games, truth-telling can be a dominant strategy. A weaker notion is that truth-telling is a Bayes-Nash equilibrium: it is best for each participant to tell the truth, provided that others are also doing so. See Harrison (2007).

[17] Their empirical analysis is an update of Foster et al. (1997).

dichotomous and multiple choice, referendum, payment card and conjoint, is important in reducing bias. There is some weak evidence that bias increases when public goods are being valued (reinforcing the evidence of List and Gallett 2001), and that some calibration methods (see below) may be effective at reducing bias. However, results are quite sensitive to model specification, which will remain a problem until a comprehensive theory of hypothetical bias is developed.

There are a number of CV studies that utilise calibration techniques to control for hypothetical bias. Studies that employ *ex ante*, or instrument calibration, techniques such as budget reminders or cheap talk scripts (Cummings and Taylor 1999, List 2001), attempt to get unbiased responses from participants. *Ex post*, or statistical calibration techniques, on the other hand, recognise that responses are biased and they attempt to control for it using laboratory experiments to calibrate field data (Fox et al. 1998) or uncertainty adjustments (Poe et al. 2005).

Blackburn et al. (1994) define a "known bias function" as one that is an *ex post* systematic statistical function of the socio-economic characteristics of the sample. If this bias is not mere noise, then one can say that it is "knowable" to a decision maker. They then test if the bias function is transferable to a distinct sample valuing a distinct good, and conclude that it is. In other words, they show that one can use the bias function estimated from one instance to calibrate the hypothetical responses in another instance, and that the calibrated hypothetical responses statistically match those observed in a paired real elicitation procedure. Johannesson et al. (1999) extend this analysis to consider responses in which subjects report *the confidence with which they would hypothetically purchase the good at the stated price*, and find that information on that confidence is a valuable predictor of hypothetical bias.

The idea of instrument calibration (first used in Harrison 2006), in contrast to statistical calibration, has generated two important innovations in the way in which hypothetical questions have been posed: recognition of some uncertainty in the subject's understanding of what a "hypothetical yes" means (Blumenschein et al. 1998, 2001), and the role of "cheap talk" scripts directly encouraging subjects to avoid hypothetical bias (Cummings and Taylor 1998, List 2001, Aadland and Caplan 2003, and Brown et al. 2003). The evidence for these procedures is mixed. Allowing for some uncertainty can allow one to adjust hypothetical responses to better match real responses. Although this could be estimable, it normally presumes that one knows *ex ante* what threshold of uncertainty is appropriate to apply (see Swait 2001). However, simply showing that there exists a threshold that can make the hypothetical responses match the real responses, once you look at the hypothetical and real responses, is not particularly useful unless that threshold provides some out-of sample predictive power. Similarly, the effects of "cheap talk" appear to be context specific, which simply means that one has to test its effect in each context rather than assume it works in all contexts[18]. There is a case to build in uncertainty *ex ante* in experimental design, given that outcomes are uncertain for many reasons (see Harrison 2006a). This recognizes that a plan is not the same as an actual action of implementation. For example, in studies investigating the role of various congestion charging schemes in contexts where they currently do not exist, introducing an attribute that attaches a probability of such a scheme actually happening is one way to control for subjective assessment shrouded in various beliefs

[18] The increasing role that in-depth interviews and focus groups are playing in the definition of choice experiments has been found by the author to add substantial credibility to the experiments. Recent studies in the context of determining the MWTP for music in gym classes, and at nightclubs and discos, which is subsequently used in Federal Court of Australia arbitrations on music royalties, confirms this.

about the reality of the offer. Greater certainty (or lesser uncertainty of an outcome) is known to influence preferences.

2.2.2 CE evidence

CE's are typically framed in a manner that adds realism, in that they closely resemble individual purchasing or use decisions. There are surprisingly few published studies that test for hypothetical bias in CE (exceptions being Alfnes and Steine 2005, Lusk 2003, Lusk and Schroeder 2004, Cameron et al. 2002, Carlsson and Martinsson 2001, List et al. 2001, Johansson-Stenman and Svedsäter (2003), Brownstone and Small 2005, and Isaccson 2007). Both Carlsson and Martinsson (2001) and Cameron et al. (2002) fail to reject a hypothesis of equal MWTP in both a real and a hypothetical setting, while Johansson-Stenman and Svedsäter (2003) reject the equality of MWTPs, and Lusk and Schroeder (2004) find that hypothetical TWTP for the good exceeds real TWTP, but fails to reject the equality of MWTPs for changes in the single attributes. Carlsson et al. (2005) also conclude that they cannot reject the hypothesis of a hypothetical bias for MWTP in choice experiments.

List et al. (2006) explore choice experiments which conveniently provide information on the purchase decision as well as the attribute value vector. The empirical work revolves around examining behaviour in two very different field settings. In the first field study, they explore hypothetical bias in the purchase decision by eliciting contributions for a threshold public good in an actual capital campaign. To extend the analysis a level deeper, in a second field experiment they examined both the purchase decision and the marginal value vector via inspection of consumption decisions in an actual marketplace. In support of CE's, both field experiments provide some evidence that hypothetical choice experiments combined with "cheap talk", be it light or heavy, can yield credible estimates of the purchase decision. *Furthermore, they find no evidence of hypothetical bias when estimating MWTP*. Yet, they do find that the "cheap talk" component might induce internal inconsistency of subjects' preferences in the choice experiment.

Lusk (2003) explored the effect of cheap talk on willingness-to-pay that was elicited via a mass mail survey (n = 4,900) for a novel food product, golden rice. Employing a double-bounded dichotomous choice question, he found that estimated WTP, calculated from hypothetical responses with cheap talk, is significantly less than willingness-to-pay estimated from hypothetical responses without cheap talk. However, consistent with List (2001), he found that cheap talk does not reduce willingness-to-pay for *experienced*, or in our case *knowledgeabl*e, consumers. For all consumers, average WTP for golden rice exceeds the price of traditional white rice. The evidence that cheap talk tends to attenuate hypothetical bias only for subjects less familiar with the good being valued by List (2001) and Lusk (2003) reinforces the importance of referencing, a key focus of the current paper (see Section 3 below).

In addition, the potential effect of 'realism' (Cummings and Taylor, 1998) or 'consequentialism' (Landry and List, 2007), or the role of 'limit cards' (Backhaus et al., 2005)[19] further supports the appeal of referencing around an experience good or alternative. The 'limit cards' approach requests the respondent to place an imaginary 'limit card' behind the stimulus he considers just sufficient to generate a choice. In this manner, the limit card combines the first preference response in CE studies with a

[19] I thank a referee for drawing this literature to my attention.

ranking position that separates acceptable stimuli from those that are not deemed capable of leading to a choice. The underlying theoretical argument to support limit cards is that individuals evaluate 'decision' alternatives at a subjective level, called the comparison level, which is not dissimilar to the idea of a reference alternative. In some sense, this literature is related to information processing which is now recognized as having a role to play in minimising hypothetical bias (see Hensher and Greene 2008). Backhaus et al., (2005) use a weekend trip to three capital cities (Paris, Rome and Vienna) as the choice context to show that the mean WTP, based on limit conjoint analysis, is very close to the real WTP; in contrast the CV mean estimates are substantially different.

The most relevant condition for choice analysis is salience, which requires that the reward be directly related to the decision the subject makes during a study. Paying a respondent a fixed amount is not salient, because there is no relationship between the respondent's performance/actions and the reward he or she receives. There is no reason to expect that the respondent's behaviour during a study will be consistent with his or her behaviour during a similar, real-world, economic activity. Ding et al. (2007) suggest that traditional marketing research often relies on conjoint analysis in which participants are either paid nothing or a non-salient reward to answer questions about hypothetical purchase decisions. Such studies struggle to uncover true consumer preferences because participants have little stake in the answers they give. They developed an approach that provides participants with incentives based on actual behaviour. Ding et al. conducted field experiments in a Chinese restaurant to test their incentive-aligned approach. Participants stated their meal preferences and eventually had to pay for and eat the preferred meal. Using the traditional, hypothetical conjoint approach they were able to predict consumer's top choice only 26 percent of the time. In contrast, using the incentive-aligned approach, they were able to predict consumer's top choice 48 percent of the time.

There are two innovative urban transport studies using choice experiments that investigate hypothetical bias. Using a simple dichotomous choice experiment with two attributes, Isacsson (2007) suggests that there is a bias in estimates of the value of time savings associated with public transport, based on hypothetical choices. Real values tend to be higher than values derived from hypothetical choices. This replicates the findings of Brownstone and Small (2005). Assuming an exponential distribution for the value of time savings, real choices in Isaccson produced an estimated mean value of time savings which was twice as large as the corresponding hypothetical value. This evidence in CE's in transportation applications, is the inverse to the general findings in CV studies that conclude that hypothetical WTP estimated in stated preference surveys is most often found to be an overstatement of true WTP (see e.g. Harrison and Rutstrom 2008, List and Gallet 2001, and Murphy et al. 2005).

2.2.3 Summary

In summary, this section has identified a number of candidate influences on the magnitude of hypothetical bias. The key influences are the use of cheap-talk to assist in attenuating bias, especially where there is a lack of experience; the ability to opt-out in contrast to a forced choice, which is linked to referencing in that the respondent's opt-out is maintenance of the status quo; and the use of processing strategies such as 'limit-cards' or questions to establish the threshold limits for the set of alternatives that are treated a serious decision alternatives.

What we appear to have is a strong recommendation for greater clarity of the choice experiment (i) in terms of a translation of offerings in real markets, (ii) in the manner in which experience is embedded into the CE (through, in particular, pivotting around an experienced good), and (iii) in the way that we capture information to delineate the process heuristics that each individual uses in evaluating the attributes and alternatives.

In the remaining sections, we focus on the role that referencing (linked to opt-out) can play in reducing hypothetical bias, taking the Brownstone and Small (2005) study as one influential and current benchmark in travel behaviour research of evidence on MWTP obtained from observing real behavioural decisions relative to CE evidence. We also acknowledge that many researchers regard the MWTP from traditional RP studies (with all their known deficiencies) as another 'benchmark of interest', which often produces higher mean estimates than CE studies. An investigation into the role of process heuristics is provided elsewhere in Hensher and Greene (2008).

3. Some background evidence in transportation studies

The general lack of evidence from real markets observed 'at a distance' in the transportation context, limits the comparisons herein to a range of empirical paradigms, ranging from the traditional RP choice data through to various choice experiment specifications. Drawing on a number of (non-referenced) data sets collected by the author over the last 20 years (summarised in Appendix A), we ran separate traditional RP and traditional (i.e. non -referenced) CE models as mixed logit error component models (see Table 1 and Appendix B). The VTTS was calculated for each RP and CE choice set and compared. These non-pivot (pencil and paper face to face survey) CE designs have attribute levels determined by a specific trip length segment, and are not individualised to each person's current or recent trip20.

Table 1: Summary of illustrative Australian empirical evidence on VTTS: Traditional CE vs. RP

Study Title	Context	Mean VTTS ($ per person hour)	Ratio RP/CE (using 95 percent confidence limits on a symmetrical distribution)
RP vs. CE evidence			
Sydney-Melbourne 1987	Long-distance non-commuting; labelled mode choice	Error component: RP: 9.74±6.23 CE: 5.81±3.01	1.67±2.1
Six Australian capital cities 1994	Urban commuting; labelled mode choice	Error component: RP:3.51±1.47 CE:4.20±2,13	0.838±0.7
Sydney Pricing Tribunal 1995	Urban commuting; labelled mode and ticket type choice	MNL RP:6.73±3.94 CE:6.11± 3.22 Error component: RP:6.87±4.58 CE: 6.26±2.95	1.10±1.2 1.09±1.56

Note: All studies used a face to face pencil and paper survey

[20] Both of these features may well be at the centre of the sources of hypothetical bias.

The evidence suggests that the ratio RP/CE is not significantly different from 1.0 in any of the three studies[21] and hence we cannot reject the null hypothesis of no evidence of hypothetical bias. If the 'truth' resided in the RP model, which in these three studies has the usual concerns with the identification of the relevant set of non-chosen alternatives (including the measurement error problem), then we would indeed conclude that hypothetical bias is not an issue.. The influential paper by Brownstone and Small (2005), especially in the USA, however, has convinced a growing number of researchers and practitioners that MWTP from CE studies are significantly under-estimated compared to *revealed behaviour* studies, and hence there is a need to seek out a possible explanation(s) for this. Brownstone and Small (2005) suggest that the traditional RP estimate is not a benchmark compared to real market observation, along the lines of Brownstone and Small (2005)[22]. Importantly Brownstone and Small's model looks like the usual RP model, but the data is obtained from a sample of traveller's actually observed choosing between a variable priced tolled lane and a free lane. The attributes are measured by external procedures so that the levels of times and costs actually experienced for both alternatives in the choice set are not subject to the usual concerns associated with asking individuals.

Given the dominance of habit (in contrast to variety seeking) in much of modal and route trip activity for a given trip purpose, there is appeal in focussing on choice situations where preference inference (and hence MWTP) might be better identified without 'forcing' non-chosen alternatives into the decision space, unless one can capture data along the lines of Brownstone and Small (see footnote 28). Our preference-revelation paradigm promotes the use of a reference (or status-quo) alternative in the design of choice experiments, which is also the opt-out under conditions of actual experience. This appears to offer great promise in the derivation of estimates of MWTP that have a closer link to the real market activity *of each individual*.

Reference alternatives have an important role to play in giving sense to the levels of the attributes offered in the CE choice scenarios. Using a computer aided personal survey instrument (CAPI) and internet based surveys, we can now automatically individualise the attribute levels in choice experiments relative to a reference experience (e.g., a recent travel activity). That is, the levels seen by each individual will differ according to the levels associated with a reference alternative, even where the design levels (as percentage variations around the reference point) are the same. What is not clear is whether the reference alternative should be included in the choice set used in model estimation. In Table 2 we investigated this issue, using a 2004 toll road study in Sydney. Respondents were asked (i) to make a choice amongst the reference alternative and two CE alternatives, and (ii) to choose amongst the CE alternatives. What we find suggests that there is a difference, albeit small, in the mean VTTS which is smaller when the error components specification is used, in contrast to the simple MNL form. The confidence intervals23 for the above values were estimated using the t-ratio method equation derived by Armstrong et al. (2001):

$$V_{S,I} = \left(\frac{\theta_t \cdot t_C}{\theta_C \cdot t_t}\right) \cdot \frac{\left(t_t t_C - \rho t^2\right)}{\left(t_C^2 - t^2\right)} \pm \left(\frac{\theta_t \cdot t_C}{\theta_C \cdot t_t}\right) \cdot \frac{\sqrt{\left(\rho t^2 - t_t t_c\right)^2 - \left(t_t^2 - t^2\right)\left(t_C^2 - t^2\right)}}{\left(t_C^2 - t^2\right)} \quad (1)$$

[21] A referee described this evidence as 'a good result'.

[22] I acknowledge personal communication with Ken Small and David Brownstone in early 2008.

[23] These are important because the estimated VTTS are ratios of random variables, so they are also random.

where t_t and t_c correspond to the *t*-ratios for parameter estimates for travel time and cost, θ_t and θ_c, respectively; t is the critical value of the statistics given the degree of confidence required and ρ is the coefficient of correlation between both parameter estimates. This expression assures positive upper and lower bounds for the VTTS if the parameters involved are statistically different from zero.

On a test of differences, the error component model findings are not statistically different at 95 percent confidence level. This test, however, says nothing about added value of the reference alternative as a way of identifying the marginal disutility of time and cost associated with an alternative chosen in an actual market setting, complete with all the real-world constraints that an individual takes into account in choosing that alternative.

Table 2: Empirical evidence on CE based VTTS (mean $ per person hour and statistical uncertainty) for Pivot Data Paradigm, treating Time and Cost Parameters Generic across all Alternatives

Study Title	Context	Model including reference alternative	Model excluding reference alternative (a form of opt out)
Toll Study Sydney 2004	Urban commuting, unlabelled choice	MNL: 18.6±6.3 Error Component 18.1±7.4	MNL: 17.85±7.7 Error Component 17.83±8.1

Constraining parameters across the reference and CE design alternatives, which is common in the majority of choice experiments, may actually be clouding real information on the difference the marginal disutility of time and cost of a real alternative and a hypothetical alternative that may be sources of differences in VTTS.

The WTP derived from choice experiments reported in Tables 1 and 2 are ratios of parameter estimates, and are typically sensitive to small changes in the numerator and/or denominator estimates, which may be differentially impacting on each alternative (although suppressed when parameters are generic across the reference and CE alternatives). In other words, deviations between RP and CE WTP estimates due to hypothetical bias might be confounded by deviations introduced by something as simple as adding another attribute to one or more alternatives.[24]. In the following sections, we need to take a closer look at the richness of the information in the numerator and denominator of a WTP calculation, and the additional information offered in a reference-based choice experiment, through separation of the utility expressions for the reference (i.e. current market decision outcome) and CE alternatives.

[24] A study by Steimetz and Brownstone (2005) cited in Brownstone and Small (2005) bootstraps the distribution of WTP and takes the mean of this distribution as a point estimate in an effort to accommodate this numerator and denominator sensitivity. I thank a referee for pointing this out.

3.1 Marginal WTP – numerator and denominator effects

Wardman (2001) and Brownstone and Small (2005) suggest a range of reasons as to why CE and RP MWTP may differ (they do not investigate TWTP25). Wardman (2001, 120) suggests that a lower CE MWTP can be explained (in part) by (i) strategic response bias, especially on the parameter of cost which appears in the denominator of the calculation of WTP in dollars, associated with greater sensitivity to cost variation that a choice experiment generates; (ii) the ability in a choice experiment to "… adopt simplified decision rules such as ignoring attributes of lesser importance or which vary less"; and (iii) a variation on (ii), to ignore attribute variations which are not realistic, thereby reducing mean parameter estimates. He suggests this is more likely to be an issue for the parameter estimate (such as travel time) which is the numerator of the WTP calculation.

Brownstone and Small (2005) also offer some explanations for the differences[26], also cited in Isacsson (2007). The most appealing is that individuals display (time) inconsistency in their actual behaviour, or more generally constraints associated with real actions that are not accounted for in choice experiments. It is suggested that these constraints tend to result in higher-cost choices more frequently in real life than in hypothetical surveys[27]. They also consider the misperception of travel time. They ask individuals to report the time savings they think could be realised by using express lanes[28]. This belief elicitation was non-incentivised. Individuals typically report an estimate (based on the mean), twice the actual time-savings. Brownstone and Small suggest two possible explanations (2005, 288): (i) individuals focus on total delays on part of the trip instead of the full origin-destination trip, and (ii) impatience with heavy traffic leads to exaggeration of actual delay time. These reasons are then used to suggest that the same level of an attribute in a choice experiment will lead to the same reaction, and hence a lowering of the parameter estimate for time.

Hensher (2006, 2008) and Hensher and Greene (2008) promote the idea of attribute processing as a behaviourally meaningful way of ensuring that individuals use heuristics that they also use in real markets (although there may be additional choice experiment-specific effects given the amount of information being offered for processing, which may not change the heuristic set but simply invoke a specific processing rule). We do not see Wardman's point (ii) as a CE-specific issue, since this happens also in RP settings. Supplementary questions should be asked to reveal such processing rules for CE *and* RP data, or model specifications defined to test for and capture specific process heuristics, up to a probability (see Hensher and Greene 2008, Hensher and Layton 2008). Furthermore Wardman's point (iii) is linked implicitly to the promotion of pivot designs (see Rose et al. 2008) that can, if carefully designed, reduce this feature of many poorly designed choice experiments (see below).

---

[25] TWTP is predominantly a focus of environmental, health, marketing and agricultural applications.

[26] Brownstone and Small (2005) suggest that the mean VTTS in a toll HOT lane vs. free route context is in the range USD$20-$40 per person hour, which is about 50 percent higher than the evidence from SP studies (i.e., $USD13-$16). The high end USD$40 is a self-selected group who had already obtained a transponder that enables them to use the express lane if they so choose, and hence they would be expected to have the highest VTTS.

[27] Higher attribute levels tend, holding unit of measurement fixed, to result in lower parameter estimates, and hence with the cost parameter in the denominator, we obtain a higher mean MWTP for RP situations compared to SP situations.

[28] David Brownstone (personal communication 22 February 2008) advises that many of the respondents in the CE RP comparisons actually switched between the tolled and untolled alternatives regularly (at least once a week). He suggests that it would be interesting to repeat the estimation on the subset of these switchers, who are quite familiar with both alternatives.

Brownstone and Small's suggestion in the context of time savings realised by using the express lane (page 288), is controversial; namely that "…if people experiencing a 10-min time delay remember it as 20 min, then they probably react to a hypothetical question involving a 20-min delay in the same way that they react to a real situation involving a 10-min delay. This would cause their measured value of time savings in the hypothetical situation to be exactly half the value observed in real situations." Unlike RP data, where one is asked to indicate the level (and in some cases the difference) or, as in the case of Brownstone and Small, use some other means of measuring not related to a specific individual's actual trip, such as floating cars and loop detectors; in a choice experiment the level is actually given to each sampled respondent. Hence an individual is processing a given level of an attribute, used in model estimation, which is not the same as asking an individual for an attribute level or obtaining it from a third-party source, for the non-chosen alternative, or the difference, and then constructing an attribute level for the non-chosen alternative. In one sense this removes an element of uncertainty associated with a respondent having to construct a level of an attribute associated with a non-chosen alternative in an RP study.

The MWTP (e.g., value of travel time savings) is shorthand for the ratio of two distinct quantities – the marginal (dis)utility of an attribute of interest (e.g., travel time), and the marginal (dis)utility of money (Hensher and Goodwin 2004). Both are confounded by changes in tastes, leisure activities, education, and opportunities or choice set open to sampled individuals, as well as the data collection paradigm (Hensher 2006, Harrison 2006, 2007). Given the different context of a choice experiment in general, it must be recognised that CE studies should be annexures to RP studies that can supplement where RP data is deficient. Pivot designs, discussed below, may well be the way forward.

3.2 Pivot designs – elements of RP and CE

RP data is generally regarded as rich in information on the chosen alternative, but problematic on the attributes describing the non-chosen alternatives. This is due, in large measure, to a lack of experience with non-chosen alternatives, common in transportation contexts, many of which exhibit strong patterns of habitual behaviour with high levels of inertia required to motivate any serious consideration of alternatives. Variety seeking is not common with many transportation choices, especially in urban settings. There will always be perceptual bias, which many would argue is not a bias, but a product of the exposure and overt experience an individual has accumulated in respect of these alternatives. There is a large literature that embodies ideas of reinforced preference towards the chosen alternative (see, for example, Tversky and Kahneman 1981, Gilboa and Schmeidler 2001). The recent development in referencing or pivot designs (see Rose et al. 2008, Hess et al. 2008) is in part a response to the observation[29] that individuals are, in many situations, habitual in behaviour, especially in the short to medium term in many transport 'choices' such as the daily commute, in contrast to variety seeking; and it often requires a perceptually different level of offering of an alternative (defined in terms of behavioural thresholds on specific attributes and/or mixes of attributes) before an individual will *consider* switching and then *possibly* switching. Given that cross-section studies, which are the typical data source for MWTP, are seen as representing long run behaviour, we have to contemplate mixtures

[29] This is generally the case with the most popular transportation application of commuter mode or route choice.

of habitual and variety seeking behavioural responses even where the bias towards the reference alternative is strong[30]. This is consistent with choice experiments offering alternatives around the reference alternative that are different labelled alternatives (e.g., public transport instead of car for the commute), in addition to the unlabelled CE alternatives where one is varying the attribute levels of the reference alternative. Conversion of consideration into action is an important threshold question.

Given the problems in identifying 'reliable' (in an ex post choosing sense) levels of attributes of non-chosen alternatives, with systematic misperception of travel times suggested by Brownstone and Small (2005, p. 288) as a possible source of differences in RP and CE MWTP, there is *potential merit* in seeing pivot-based choice experiments as richer in reality than traditional RP with forced (and often artificially constructed) alternatives. A CE-pivot experiment (CE_PV) recognizes that an individual has chosen an alternative in a real market setting and that this alternative is reasonably assumed to be the utility maximising outcome (given all the perceptions and awareness that exist in actual markets)[31]. An individual when asked to evaluate a well designed labelled or unlabelled CE_PV, is in relatively familiar territory compared to an RP context where attribute levels of non-chosen alternatives are typically unknown and 'wild guesses' are not uncommon. Specifically a CE_PV experiment offers design variations around an experienced alternative, which are selected to ensure the attributes and their mix are comprehensiveness and comprehendible (Hensher 2006). Pivoting is one way of promoting *relevancy* in attribute levels, in line with ideas in prospect theory (Tversky and Kahneman 1981).

Hensher (2001) demonstrates that CE-based WTP estimates depend largely on the *context* of the alternatives presented, such as "start/stop" traffic versus "free-flow" traffic (as Wardman (2001) does, but to a lesser extent). This illustrates the importance of pivoting CE alternatives off of real world experience – enabling respondents to apply some meaningful context to these hypothetical alternatives. Second, the RP-based findings of Hensher (2006a) suggest that VTTS estimates can reflect additional congestion costs (a finding confirmed by Steimetz 2008). As such, a possible explanation for the divergence between RP and CE estimates is that CE respondents cannot visualise all of the congestion costs that would, in the real world, accompany the hypothetical travel times presented to them. But pivoting off of a real-world alternative might enable respondents to visualise such costs, offering some promise of tightening the gap between RP and CE results.

To investigate the role of referencing around an experienced alternative (essentially an RP observation in an actual market), we re-estimated some models that were developed for value of travel time savings studies in the context of tolled vs. non-tolled routes in Australia and New Zealand. We focus on (habitual) commuting activity in the Sydney region in 2004 and in a regional context in New Zealand in 2007. A typical CE screen is shown in Figure 1 (taken from the Sydney study), extracted from a computer aided personal survey instrument.

---

[30] I acknowledge discussions with Ken Small on this point.

[31] Hensher at al. (2005) have suggested that one might estimate stand-alone CE models to obtain robust parameter estimates on each attribute and then calibrate the constants to reproduce base market shares observed in real markets. This removes the need to estimate RP models. This approach is conditional on assuming that the parameter estimates obtained from RP alternatives, be they from a stand-alone RP model or a joint (rescaled) RP-CE model, are statistically and behaviourally less reliable.than from CE alternatives, and especially the reference alternative..

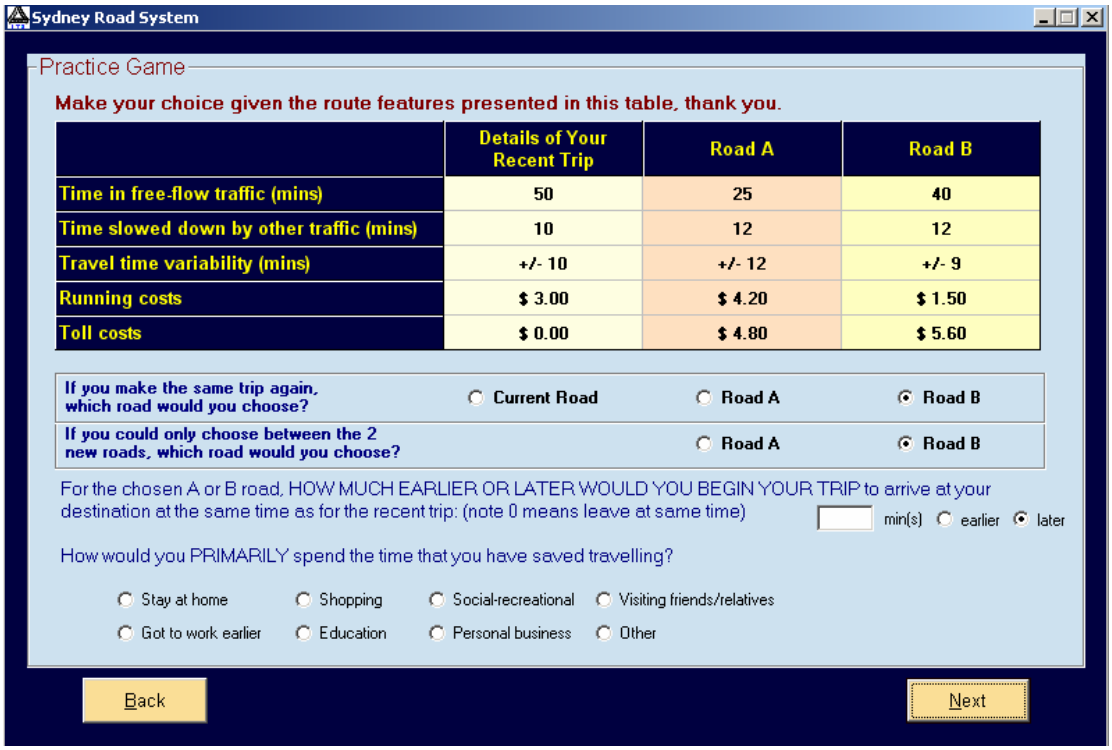

Figure 1: An illustrative SC screen from a CAPI

The initial Sydney model estimation involved estimation of a choice between the reference alternative, defined in terms of a recent or current commuting trip, and two CE alternatives, the latter developed as a D-efficient design (see Rose et al. 2008), with actual attribute levels pivoted around the reference alternative's levels. The subsequent Sydney model focused on the choice between the two CE alternatives (essentially a forced choice). The New Zealand study involved the exact same model specification. Mixed logit models with error components (to account for scale)[32] were estimated, with the parameters for travel time for each of the reference and CE alternatives specified as random with triangular distributions[33]. Separate fixed parameters were estimated for cost (running and toll cost as applicable)[34]. We also included two constants to account for any biases in favour of the reference alternative and the first ordered CE alternative.

The key findings are summarised in Table 3.[35] The models in Table 3 had the best overall goodness-of-fit (on a likelihood ratio test). The VTTS estimates are conditional estimates based on the full distributions, and not the means. For the Sydney study, the mean of the VTTS distribution for the reference alternative is $26.99 per person hour with a standard deviation of $7.94; the mean for the CE alternatives is $17.92 (standard deviation of $7.82), derived from the model that includes the reference alternative. The

[32] In part to recognise the greater uncertainty about the SC designed alternatives relative to the reference alternative for each respondent.

[33] The models used simulated MLE with 500 Halton draws and accounted for the correlation between 16 choice scenarios shown to each sampled respondent.

[34] A referee indicated that Sillano and Ortúzar (2005) argue that: "constraining a taste coefficient to be fixed over the population, may make it grow in a less than average proportion (i.e. the parameters that are allowed to vary grow more than the parameters that should vary over the population, but are constrained to be fixed)".If this is the case then it would apply to both the reference and CE alternatives. In addition, the majority of empirical RP studies using mixed logit also impose this condition.

[35] This included unconstrained triangular and normal distributions for travel time, and random parameters specifications for travel cost. The MNL model had a significantly worse overall fit (see footnote to Table 3), and produced ratios of RP:CE mean VTTS of 1.46 and 1.05 respectively for Sydney and New Zealand.

forced choice models produced a mean VTTS of $23.24 per person hour (with a standard deviation of $7.52). The ratio of the Reference to CE alternatives mean VTTS is 1.51. For the New Zealand study, the mean VTTS for the reference alternative is $27.34 per person hour, with a standard deviation of $7.46; the mean for the CE alternatives is $13.65 (standard deviation of $4.31), derived from the model that includes the reference alternative. The forced choice models gave a mean VTTS of $11.28 per person hour (with a standard deviation of $5.35). The ratio of the Reference to CE alternatives mean VTTS is 2.00. A t-ratio test of differences shows that the WTP associated with the reference alternative and the CE alternative are statistically significant at the 95 percent confidence level.

Table 3:. Summary of findings for pivot-based models m

Mean parameter estimates (for times and costs only)[36]:

Study	Reference alternative		CE alternatives	
	Time	Cost	Time	Cost
Sydney	-0.1008 (-7.7)	-0.2239 (-12.049)	-0.0669 (-16.6)	-0.2138 (-11.3)
New Zealand	-0.2128 (-9.25)	-0.4774 (-2.1)	-0.1783 (-16.80)	-0.7634 (-32.9)

WTP estimates:

Study	Reference alternative	CE alternatives	Forced choice (CE only)	Ratio Ref:CE
	Mean (standard deviation)			
Sydney	26.99(7.94)	17.92 (7.82)	23.24 (7.52)	1.51
New Zealand	27.34 (7.46)	13.65 (4.31)	11.28 (5.35)	2.00

Note: log-likelihood for Sydney (912 observations) and New Zealand (1840 observations) models are respectively -662.51 and -1187.96 (Multinomial logit log-likelihoods are respectively -837.8 and -1630.2).

We find that the marginal disutility associated with travel time in the reference alternative is substantially higher (especially for Sydney) than that associated with the CE design alternatives, and is either similar (i.e., Sydney) or lower (i.e. New Zealand) for cost; resulting in the higher mean VTTS for the reference (or real market) alternative. The evidence from other studies by Hensher and Louviere (see Hensher 2006, Louviere and Hensher 2001) that the attribute range has the greatest influence on MWTP than any other dimension of choice experiments,[37] with MWTP being higher with a reduced attribute range, supports the findings herein; the CE design alternatives have a wider attribute range relative to the range of attributes of other alternatives that people face in real choices, and hence a lower mean VTTS than the mean VTTS from the real market alternative. If we take the Sydney sample as an example, the ratio of the range of each attribute in the numerator and denominator of the calculation of VTTS for the reference and CE alternatives is 1.42 for time and 1.48 for cost. The ratio of the reference alternative to CE VTTS is 1.51; hence are we seeing a coincidence or something of empirical interest as a statistical calibration (ex post) adjustment to 'explain' the difference between the VTTS?

[36] We do not report the reference-alternative specific constant and the stated choice dummy variable for choice scenario 1, both of which account for the mean influence of other attributes and context.

37 Hensher and Louviere have found, in many studies, that the MWTP increases as the range of attribute levels decreases and vice versa. In CE studies it is common to have a wider range of an attribute to assess; that is essentially what choice experiments are all about, creating behaviourally richer variance. However this may come at a price, in that real markets are not so rich in variability, and hence when actual market data is used, we observe after estimation, higher MWTP compared to a stated choice experiment. This naturally begs the question – does the ratio of the range of each attribute in the numerator and denominator of the calculation of MWTP for the Reference and CE observations account for part or all of the difference in the mean MWTP?

To comment further on the influence of attribute range, which has found to be the major dimension of a CE influencing WTP, research in marketing (e.g., Ohler, et al, 2000) suggests that heterogeneity systematically varies with attribute range and distribution, as do model alternative-specific constants and goodness-of-fit measures (see also McClelland and Judd 1993), but preference model parameters remain largely unaffected. Thus, it is unclear what to make of empirical heterogeneity results because they may prove to be largely contextual; that is, they are associated with particular patterns of attribute ranges and samples of people, and cannot be generalised without taking differences in attribute ranges and people into account. The need to take into account links with characteristics of choosers and heterogeneity distributions has been recognised (see Hensher 2006), but there has been little recognition of the fact that if one changes the range and/or distribution of attributes in design matrices, this can lead to significant differences in inferences about heterogeneity. Simply put, the greater relevance in preserving the attribute content under a wider range will mean that such an attribute is relatively more important to the outcome, than it is under a narrow range specification, and hence a higher mean WTP is inferred (Louviere and Hensher 2001).

The empirical evidence on VTTS from the two studies is in line with the relative magnitudes of SC and RP mean MWTP found by Brownstone and Small (2005)[38] as long as we accept that under habitual behaviour the reference alternative has important information on the marginal disutility of attribute levels associated with the experienced alternative. The difference between our studies and those of Brownstone and Small, is that we focused on a known trip, and assumed that most commuters have little idea about the non-chosen alternative(s). The latter, one might argue, in an RP setting, exists to enable the estimation of a choice model, and to give variability in trip attributes. Under conditions of habitual behaviour, a well designed pivot-based choice experiment can deliver the relevant market information as well as attribute variability, while avoiding the problems in identifying meaningful data on non-chosen alternatives, especially in contexts where habit and inertia are very strong elements of real market behaviour. The findings support the relative magnitudes of MWTP found by Brownstone and Small (2005) and Isacsson (2007). If one desires to use traditional RP MWTP as the benchmark, which in the non-transport literature suggests the MWTP from CE studies are on the low side, then the findings herein are consistent with closing the gap on hypothetical bias. If RP and CE studies in transport cannot establish any evidence on hypothetical bias, then one wonders why we have invested so much in choice experiments[39].

The appeal of pivoting is not to imply specifying time and cost parameters as generic across all alternatives, but to recognise the role of CE data in generating variability about the real market experience (i.e., the pivot) in order to be able to estimate parameters. The argument is that this looks like offering a richer attribute preference revelation setting than either (i) the current view on RP, with problematic identification of non-chosen alternatives, and (ii) the treatment of the CE alternatives as having 'equal' status as the pivot alternative is real market identification. Crucially, however, we need the CE alternatives (without measurement error, but subject to respondent perception), to provide the necessary variation in attribute data to reveal preferences.

[38] Given the 2004 exchange rate of AUD$1=$USD$0.689, the Sydney evidence for the reference alternative is USD$39.48; compared to the SC estimate of (i) USD$19.93 for the model that includes the Reference alternative, or (ii) USD$16.08 when the forced choice amongst two CE alternatives is used.

[39] Except where the focus is on new alternatives and possibly very large attribute changes associated with existing alternatives that are outside the range of market experience.

The support for this approach is in part reinforced by the evidence from Brownstone and Small (2005) and Isacsson (2007) on the relativity of the market WTP against the CE evidence from studies where actual trade-offs are being observed and measured in real markets.

The empirical evidence herein suggests that, for all the years of interest in choice experiments, and the debate about the role of traditional RP and CE data, we may have missed or masked an important message; namely that choice experiments with referencing back to a real market activity, especially where it is chosen on repeated occasions, may provide a suitable specification, short of capturing data 'at a distance', where the latter has evaded every single travel study to date[40]. If we recognise that the requirement to seek data on at least one non-chosen alternative in RP modelling is linked to the creation of variance necessary to estimate a model, then this imposition in the context of habitual behaviour may be accommodated by variance revelation through an CE pivot design, where the only information required from real markets relates to the habitually selected alternative.

We strongly recommend further research into the proposition that future choice experiments should consider using a real market reference alternative as a pivot in the design of the choice scenarios[41]. This not only grounds the experiment in reality *at an individual respondent level*, it also enables estimation of alternative-specific parameters for relevant attributes that enable derivation of estimates of MWTP for the real market alternative, and separately, for the CE alternatives. Pivot based CE data has the power of richness to enable respondents to express preferences involving not only the actual memory but also *related hypothetical memories constructed from it* (Hensher 2006). We do, however, emphasise that the evidence herein in support of the directional and magnitude differences between WTP associated with RP and CE alternatives should not be seen as anything more than encouraging consistency and hence reducing the gap in respect of hypothetical bias. Natural field experiments are required to test this preliminary finding.

4. Conclusions

This paper has brought together elements of the literature on revealed and stated choice studies (CV and CE) to identify the nature and extent of hypothetical bias, and what might be sensible specifications of data and models to reduce the gap between MWTP estimates likely to exist in actual markets, when observed 'at a distance', and estimates from choice experiments.

In suggesting that mean MWTP for time savings is lower when trading time and cost in utility expressions associated with SC alternatives compared to RP alternatives, we

[40] Brownstone and Small measured travel times of each alternative with floating cars (on SR91) and loop detectors on I-15 which is the closest we have come to real independent observation.

[41] This should include, or at least consider, the development of models in which we can account for sign-dependent preferences with respect to a reference point outcome (e.g., Hess et al. 2008), as suggested by cumulative prospect theory (CPT). Seror (2007), in the context of women's choices about prenatal diagnosis of down syndrome concluded that CPT fitted the observed choices better than expected utility theory and rank-dependent utility theory. Such a finding has been questioned by a number of researchers claiming that many studies have been far too casual about what "the" reference point is, and allowed their priors, that loss aversion is significant to drive their specification of the reference point. See Andersen et al. (2007a). In general the notion of a reference point makes good sense in typical transport applications.

recognise that there is limited (but powerful) evidence promoting this relativity from the very influential paper by Brownstone and Small (2005)[42], reinforced by Isacsson (2007). A way forward within the context of choice experiments, when the interest is on estimating MWTP under conditions of habit, which is common in many transport applications, is to recognise the real market information present in a reference alternative. What we find, empirically, is that when a pivoted design is used for constructing choice experiments, and the model is specified to have estimated parameters of time and cost that are different for the reference alternative than the hypothetical alternatives, the estimated value of travel time savings is higher for the reference alternative than for the hypothetical alternatives. This model specification is not the specification that researchers have generally used with data from pivoted experimental designs. Usually, time and cost are specified to have the same parameters for the reference and hypothetical alternatives. The proposal herein for reducing hypothetical bias (given the Brownstone-Small 'benchmark'), is to use a pivoted design and allow different parameters for the reference and hypothetical alternatives.

Despite the importance of good experimental design, the disproportionate amount of focus in recent years on the actual design of the choice experiment, in terms of its statistical properties, may be at the expense of substantially placing less focus on real behavioural influences on outcomes that require a more considered assessment of process (see Hensher 2008), especially referencing that is grounded in reality.

There are many suggestions from the literature, derived from mixtures of empirical evidence, carefully argued theoretical and behavioural positions, and speculative explanation. The main points to emerge, that appear to offer sensible directions for specifications of future choice studies, are:

1. The inclusion of a well-scripted presentation (including cheap talk scripts), explaining the objectives of the choice experiment;
2. Inclusion of the opt-out or null alternative, avoiding a forced choice setting unless an opt-out is not sensible;
3. Pivoting the attribute levels of a choice experiment around a reference alternative that has been experienced, and/or there is substantial awareness of, and estimating unique parameter estimates for the reference alternative, in order to calculate estimates of marginal willingness to pay for an alternative that is actually chosen in a real market;
4. The ability to calibrate the alternative-specific constants through choice-based weights on alternatives where actual shares are known. This may not be feasible in many applications, but where there is evidence of actual market shares on the same alternatives, this is essential if a valid comparison of TWTP is to be made[43].
5. The inclusion of supplementary questions designed to identify the attribute processing strategy adopted, as well as a question to establish '*the confidence with which an individual would hypothetically purchase or use the good (or alternative) that is actually chosen in the choice experiment'*; the latter possibly being added into the choice experiment after each choice scenario and after an

[42] The Brownstone and Small paper is increasingly being referenced by bankers engaged in toll road project financing.

[43] Where the data relates to labelled alternatives (e.g., specific routes or modes), the pooling of data across individuals, who each evaluated the attribute packages around their chosen alternative, enables construction of a choice model that looks like the traditional RP model form. This can then be calibrated with choice-based weights.

additional response in the form of a rating of the alternatives, possible along the lines of 'limit-cards'.[44] Fuji and Garling (2003) offer some ideas on the certainty scale question.

6. Identifying constraints that may impact on actual choices that might be ignored in choice experiments, which encourage responses without commitment. Once identified, these constraints should be used in revising choice responses. How this might be defined is a challenge for ongoing research.

We also support future empirical studies that can confirm or deny the growing body of evidence on hypothetical bias in choice experiments. Using a tollroad context as an example, an empirical study might be undertaken of the following form:

1. The context is the choice amongst competing existing tolled and non-tolled routes including the option to consider none-of-these.
2. The attributes of interest should be, as a minimum, door-to-door travel time and cost, where the latter is running cost and toll cost for the tolled route, and running cost for the non-tolled route.
3. The sampled individuals are persons who currently use one of the two routes. This defines a reference alternative.
4. There are two groups:
   a. *Group A* participate in a stated choice experiment with no endowment and no randomly selected alternative for implementation, as is often practice in CV studies.
   b. *Group B* is given an endowment (e.g., a $20 subsidy voucher) and told that the voucher is a subsidy towards the toll on any tolled route, which is valid up to two weeks. The money is not a reward for participation. This is common practice in many CV and dichotomous choice studies in environmental and agricultural applications.

   We have selected the two groups as a way to test some of the imposed conditions common in many of the studies outside of transportation, as reported in this paper.
5. For each choice scenario, the sampled individual is asked to choose between (i) the reference alternative, two design alternatives, and an opt-out alternative, (ii) the reference alternative and two design alternatives, (iii) the two design alternatives and an opt-out alternative, and (iv) the two design alternatives.
6. Where the travel time is earlier or later than what one normally travels, we should identify the extent to which the individual is able to adjust their commitments to commence and/or finish the trip. This is a way of attempting to identify schedule inconvenience raised by Brownstone and Small (2005) as one reason for divergence between RP and CE VTTS.
7. A supplementary *certainty scale* question after each choice scenario, along lines suggested by Johannesson et al. (1999), on a scale 0 (very unsure) to 10 (very sure), to indicate how sure or certain the respondent is that they would actually chose that route (or not at all) at the indicated price and travel time.

[44] This, and the limit-card notion, is being tested in current research by Hensher, Rose and Beck in the context of the purchase of a new car.

Appendix A: Background to the data sets reported in Tables 1 and 2

Sydney-Melbourne 1987

The 1987 data was collected as part of a pre-feasibility study for a proposed very fast train between Sydney and Melbourne, a distance of 850 kms. An orthogonal main effects design was used to define attribute combinations for existing modes (car, coach, plane, conventional train), and the new fast train, based on three trip length segments. The RP choice set included all four existing alternatives, giving nine alternatives per respondent. There were four choice scenarios in the choice experiment. The attributes in the RP and CE data for each mode each had three levels and are in-vehicle time and cost for main mode, access time and cost, and transfer time for public modes.

Six Australian capital cities 1994

The 1994 data was collected as part of a larger study to develop an integrated transport, land use and environment impact simulator (see Hensher 2002). The mode choice context was commuting in six capital cities in Australia (Sydney, Melbourne, Adelaide, Brisbane, Darwin and Canberra), with the RP data defined as the chosen plus one alternative. Four alternatives appear in each stated choice scenario: car no toll, car toll road, bus or busway and train or light rail. The five attributes for the public transport alternatives are total in vehicle time, frequency of service, closest stop to home, closest stop to destination, and fare. The attributes for the car alternatives are travel time, fuel cost, parking cost, travel time variability, and for the toll road departure time and toll charge. Three levels were selected for each attribute. The design allows for six alternative-specific main effect models for car no toll, car toll road, bus, busway, train, and light rail. Linear by linear interactions are estimable for both car models, and generically for the bus/busway and train/light rail models. While cross effects have been assumed negligible, the four-alternative design is perfectly balanced across all attributes. The design was blocked into 27 versions of size three, and balanced for every attribute; that is each person sees each level of each attribute exactly once. The base design for the travel choice task was a 27 x 327 orthogonal fractional factorial in 81 runs. The 27-level factor was used for creating the versions, and the three level factors selected and arranged to meet the design requirements. Two factors each at two levels were constructed for the bus/busway and train/light rail modes. In the design, bus and train appear in 36 scenarios, while busway and light rail appear in 45 scanarios.

Sydney Pricing Tribunal 1995

A survey of a sample of commuters and non-commuters was undertaken in the Sydney Metropolitan Area in 1995 as part of an inquiry into the mix and level of public transport fares. Within each market segment, patterns of modal and ticket use behaviour were captured to identify both current behaviour, and the potential to switch to alternative modal and ticket use behaviour under a range of alternative fares policies for the government bus, ferry and train systems. In the survey, respondents were asked to think about the last commuter trip they made, where they went, how they travelled, how much it cost etc.; then they were asked to describe another way they could have made that trip if their current mode was not available. The current behaviour provides the revealed preference data. The stated preference component of the survey varied public

transport fares of current and alternative methods of travel under a series of different pricing scenarios. The choice set was determined exogenously, based on the physical availability of each alternative (including the availability of a car as a driver or passenger) for the journey to work. Ticket prices were varied from current levels to 50 percent above and below current levels. Each respondent was presented with four replications or scenarios for the available choice set. Three fractional factorial designs were developed for bus vs. train (eight ticket types), bus vs. car (four ticket types and car), and train vs. car (four ticket types and car). The choice response identifies the mode of transport and, for public transport, the fare they would use. Automobile operating cost was set at the marginal perceived cost of nine cents/km. See Hensher (1998) for more details.

Toll Study Sydney 2004

The data is drawn from a study undertaken in Sydney in 2004, in the context of car driving commuters making choices from a range of level of service packages defined in terms of travel times and costs, including a toll where applicable. The choice experiment presented respondents with sixteen choice situations, each giving a choice between their current (reference) route and two alternative routes with varying trip attributes. The sample of 243 effective interviews, each responding to 16 choice sets, resulted in 3,888 observations for model estimation. To ensure that we captured a large number of travel circumstances, and potential attribute processing rules, we sampled individuals who had recently undertaken trips of various travel times, in locations where tollroads currently exist. To ensure some variety in trip length, three segments were investigated: no more than 30 minutes, 31 to 60 minutes, and more than 61 minutes (capped at two hours). A telephone call was used to establish eligible participants from households stratified geographically, and a time and location agreed for a face-to-face computer aided personal interview (CAPI). A statistically efficient design (see Rose and Bliemer 2007) that is pivotted around the knowledge base of travellers is used to establish the attribute packages in each choice scenario, in recognition of supporting theories in behavioural and cognitive psychology and economics, such as prospect theory. The two stated choice alternatives are unlabelled routes. The trip attributes associated with each route are free flow time, slowed down tine, trip time variability, running cost and toll cost. All attributes of the choice experiment alternatives are based on the values of the current trip. Variability in travel time for the current alternative was calculated as the difference between the longest and shortest trip time provided in non-CE questions. The CE alternative values for this attribute are variations around the total trip time. For all other attributes, the values for the CE alternatives are variations around the values for the current trip. The variations used for each attribute are given in Table A1. Further details of the design of the choice experiment are provided in Hensher and Layton (2008).

Table A1: Profile of the Attribute range in the CE design

	Free-flow time	Slowed down time	Variability	Running costs	Toll costs
Level 1	- 50%	- 50%	+ 5%	- 50%	- 100%
Level 2	- 20%	- 20%	+ 10%	- 20%	+ 20%
Level 3	+ 10%	+ 10%	+ 15%	+ 10%	+ 40%
Level 4	+ 40%	+ 40%	+ 20%	+ 40%	+ 60%

The experimental design has one version of 16 choice sets (games). The design has no dominance given the assumptions that less of all attributes is better.[45] The distinction between free flow and slowed down time is designed to promote the differences in the quality of travel time between various routes – especially a tolled route and a non-tolled route, and is separate to the influence of total time. Free flow time is interpreted with reference to a trip at 3 am in the morning when there are no delays due to traffic.[46]

Appendix B: Summary of mixed/logit with error components

The (relative) utility of alternative *j* for individual *i* can be written, assuming linear in parameters, as:

$$U_{ri} = \alpha_{ri} + \beta_{ri}\mathbf{X_{ri}} + \varepsilon_{ri}$$
$$U_{ji} = \alpha_{ji} + \beta_{ji}\mathbf{X_{ji}} + \varepsilon_{ji} \quad (2)$$

where α_{ri} (α_{ji}) is an alternative-specific constant for the reference alternative *r* (SC alternative *j*) and individual *i*; $\mathbf{X_{ri}}$ ($\mathbf{X_{ji}}$)is a vector of attributes associated with alternative *r* (or *j*) for individual *i*; β_{ri} (β_{ji})is a vector of parameters; and ε_{ri} (ε_{ji}) is a random component that captures through a series of assumptions (see below) the unobserved sources of preference heterogeneity that can be ascribed to attributes and alternatives *r* and *j*. Within the mixed logit framework, random taste heterogeneity can be aligned to attributes through random parameters and to alternatives through error components.

The mixed logit model with all components in choice setting t is given in (3) (see Greene and Hensher 2007).

$$\text{Prob}(y_{it}=j) = \frac{\exp\left[\alpha_{ji} + \boldsymbol{\beta}_i'\mathbf{x}_{jit} + \boldsymbol{\Sigma}_{m=1}^{M} d_{jm}\theta_m E_{im}\right]}{\sum_{q=1}^{J_i}\exp\left[\alpha_{qi} + \boldsymbol{\beta}_i'\mathbf{x}_{qit} + \boldsymbol{\Sigma}_{m=1}^{M} d_{qm}\theta_m E_{im}\right]} \quad (3)$$

$(\alpha_{ji},\boldsymbol{\beta}_i) = (\alpha_j,\boldsymbol{\beta}) + \boldsymbol{\Gamma}\boldsymbol{\Omega}_i\mathbf{v}_i$ are random alternative-specific constants and taste parameters; $\boldsymbol{\Omega}_i$ = diag(σ_1,...,σ_k); and $\boldsymbol{\beta}$,α_{ji} are constant terms in the distributions of the random taste parameters. Uncorrelated parameters with homogeneous means and variances are defined by $\beta_{ik} = \beta_k + \sigma_k v_{ik}$ when $\boldsymbol{\Gamma} = \mathbf{I}$, $\boldsymbol{\Omega}_i$ = diag(σ_1,...,σ_k), $\mathbf{x}_{jit}$ are observed choice attributes and individual characteristics, and $\mathbf{v}_i$ is random unobserved taste variation, with mean vector **0** and covariance matrix **I.** This model accommodates correlated parameters with homogeneous means through defining $\beta_{i\,k} = \beta_k + \boldsymbol{\Sigma}_{s=1}^{k}\Gamma_{ks}\,v_{is}$ when $\boldsymbol{\Gamma} \neq \mathbf{I}$, and $\boldsymbol{\Omega}_i$ = diag(σ_1,...,σ_k), with $\boldsymbol{\Gamma}$ defined as a lower triangular matrix with ones on the diagonal that allows correlation across random parameters when $\boldsymbol{\Gamma} \neq \mathbf{I}$. An additional layer of individual heterogeneity can be added to the model in the form of the

[45] The survey designs are available from http://www.itls.usyd.edu.au/about_itls/staff/johnr.asp.

[46] This distinction does not imply that there is a specific minute of a trip that is free flow per se but it does tell respondents that there is a certain amount of the total time that is slowed down due to traffic etc and hence a balance is not slowed down (i.e., is free flow like one observes typically at 3am in the morning).

error components. The individual specific underlying random error components are introduced through the term E_{im}, $m = 1,...,M$, $E_{im} \sim N[0,1]$, given $d_{jm} = 1$ if E_{im} appears in utility for alternative j and 0 otherwise, and θ_m is a dispersion factor for error component m.

References

Aadland, D. and Caplan, A.J. (2003) Willingness to pay for curbside recycling with detection and mitigation of hypothetical bias, *American Journal of Agricultural Economics*, 85, 492-502.

Alfnes, F. and Steine, G. (2005) None-of-these bias in hypothetical choice experiments, *Discussion Paper DP-06/05*, Department of Economics and Resources Management, Norwegian University of Life Sciences, Aas.

Alfnes, F., Guttormsen, A. Steine, G. and 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.

Anderson, S., Harrison, G., Lau, M. and Rutström, E. (2007) Valuation using multiple price list formats, *Applied Economics*, 39, 675-682.

Anderson, S., Harrison, G., Lau, M. and Rutström, E. (2007a) Dual criteria decisions, *Working Paper 06-11*, Department of Economics, College of Business Administration, University of Central Florida.

Armstrong, P.M., Garrido, R.A. and Ortúzar, J. de D. (2001) Confidence intervals to bound the value of time, *Transportation Research* 37E, 143-161.

Backhaus, K., R. Wilken, M. Voeth and C. Sichtmann (2005) An empirical comparison of methods to measure willingness to pay by examining the hypothetical bias, *International Journal of Market Research* 47 (5), 543-562

Ben-Akiva, M.E., Bradley, M., Morikawa, T., Benjamin, J., Novak, T.P., Oppewal, H. & Rao, V. (1994) Combining revealed and stated preferences data. *Marketing Letters*, 5 (4), 336–350.

Blackburn, M., G.W. Harrison, and E.E. Rutström (1994) Statistical bias functions and informative hypothetical surveys, *American Journal of Agricultural Economics*, 76(5), 1084-1088.

Blumenschein, K., M. Johannesson, G.C. Blomquist, B. Liljas, and R.M. O'Coner (1998) Experimental results on expressed certainty and hypothetical bias in contingent valuation, *Southern Economic Journal*, 65, 169-177.

Blumenschein, K., M. Johanneson, K.K Yokoyama and P.R. Freeman (2001) Hypothetical versus real willingness to pay in the health care sector: results from a field experiment, *Journal of Health Economics*, 20, 441-457.

Brown, T.C., I. Ajzen and D. Hrubes (2003) Further tests of entreaties to avoid hypothetical bias in referendum contingent valuation, *Journal of Environmental Economics and Management*, 46(2), 353-361.

Brownstone, D. and Small. K. (2005) Valuing time and reliability: assessing the evidence from road pricing demonstrations, *Transportation Research* , 39A, 279-293.

Cameron, T.A., G.L. Poe, R.G. Ethier and W.D. Schulze (2002) Alternative non-market value- elicitation methods: are the underlying preferences the same? *Journal of Environmental Economics and Management*, 44, 391-425.

Carlsson, F and Martinsson, P. (2001) Do hypothetical and actual marginal willingness to pay differ in choice experiments? *Journal of Environmental Economics and Management* 41, 179-192.

Carlsson,F., Frykblom, P. and Lagerkvist, C-J. (2005) Using cheap- talk as a test of validity of choice experiments, *Economics Letters* 89, 147-152.

Carson, R., Flores, E. Martin, K. and Wright, J. (1996) Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods, *Land Economics*, 72, 80-99.

Cummings, R.G. and Taylor, L.O. (1998) Does realism matter in contingent valuation surveys? *Land Economics*, 74(2), 203-215.

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

Cummings, R.G., G.W. Harrison and E.E. Rutström (1995) Homegrown values and hypothetical surveys: is the dichotomous choice approach incentive compatible?" *American Economic Review*, 85(1), 260-266.

Cummings, R.G., G.W. Harrison and L.L. Osborne (1995a) Can the bias of contingent valuation be reduced? evidence from the laboratory, Economics Working Paper B-95-03, Division of Research, College of Business Administration, University of South Carolina, (see http://www.bus.ucf.edu/gharrison/wp/).

Diamond, P. and Hausman, J. (1994) Contingent valuation: is some number better than no number, *Journal of Economic Perspectives*, 8(4), 45-64.

Ding, M., Grewal, R. and Liechty, J. (2007) An incentive-aligned mechanism for conjoint analysis, *Journal of Marketing Research* XLIV (May 2007), 214-223.

Foster, V., Bateman, I. and Harley, D. (1997) Real and hypothetical willingness to pay for environmental preservation: a non-experimental comparison, *Journal of Agricultural Economics*, 48, 123-138.

Fox, J.A., J.F. Shogren, D.J. Hayes, and J.B. Kliebenstein, (1998) CVM-X: Calibrating contingent values with experimental auction markets, *American Journal of Agricultural Economics,* 80, 455-465.

Frykblom, P. (1997) Hypothetical question modes and real willingness to pay, *Journal of Environmental Economics and Management,* 34, 274-287.

Fujii, S. and T. Gärling (2003) Application of attitude theory for improved predictive accuracy of stated preference methods in travel demand analysis, *Transportation Research,* 37A, 389-402.

Ghosh, A. (2001) Valuing time and reliability: commuters_ mode choice from a real time congestion pricing experiment. Ph.D. Dissertation, Department of Economics, University of California at Irvine.

Gilboa, I. and D. Schmeidler (2001) *A Theory of Case-Based Decisions*, Cambridge University Press, Cambridge.

Greene, W. H. and Hensher, D.A. (2007) Heteroscedastic control for random coefficients and error components in mixed logit' *Transportation Research,* 43E, 610-23.

Hanemann, W. (1994) Valuing the environment through contingent valuation, *Journal of Economic Perspectives*, 8(4), 19-43.

.Harrison, G. W. (2006) Experimental evidence of alternative environmental valuation methods, *Environmental and Resource Economics,* 34, 125-162.

Harrison, G.W. (2006a) Hypothetical bias over uncertain outcomes, in List, J.A. (ed.) *Using Experimental Methods in Environmental and Resource Economics*, Edward Elgar, Northampton, MA, 41-69.

Harrison, G. W. (2007) Making choice studies incentive compatible, in Kanninen, B. (ed.) *Valuing Environmental Amenities Using Stated Choice Studies*, Springer, The Netherlands, 67-110.

Harrison, G. W., and J.A. List (2004) Field experiments, *Journal of Economic Literature*, 42(4), 1013-1059.

Harrison, G.W. and E.E. Rutström (2008) Experimental evidence on the existence of hypothetical bias in value elicitation methods, in *Handbook of Experimental Economics Results*, C.R. Plott and V.L.Smith, Eds., Amsterdam: North-Holland.

Hensher, D. (1998) Establishing a fare elasticity regime for urban passenger transport, *Journal of Transport Economics and Policy*, 32 (2), 221-246.

Hensher, D.A. (2001) Measurement of the valuation of travel time savings, *Journal of Transport Economics and Policy* 35(1),71–98.

Hensher, D.A. (2002) A systematic assessment of the environmental impacts of transport policy: an end use perspective, *Environmental and Resource Economics*. 22(1-2), 185-217

Hensher, D.A. (2006) How do respondents handle stated choice experiments? - attribute processing strategies under varying information load, *Journal of Applied Econometrics,* 21, 861-878

Hensher, D.A. (2006a) Integrating accident and travel delay externalities in an urban context, *Transport Reviews*, 26 (4), 521-534.

Hensher, D.A. (2008) Joint estimation of process and outcome in choice experiments and implications for willingness to pay, *Journal of Transport Economics and Policy*, 42 (2), May, 297-322.

Hensher, D.A. and Goodwin, P.B. (2004) Implementation of values of time savings: the extended set of considerations in a tollroad context, *Transport Policy* 11 (2), 171-181.

Hensher, D.A. and Greene, W.H. (2008) Non-attendance and dual processing of common-metric attributes in choice analysis: a latent class specification, Institute of Transport and Logistics Studies, August.

Hensher, D.A. and Layton, D. (2008) Common-metric attribute parameter transfer and cognitive rationalisation: implications for willingness to pay, Institute of Transport and Logistics Studies, July.

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

Hess, S., Rose, J. M. and Hensher, D.A. (2008) Asymmetrical preference formation in willingness to pay estimates in discrete choice models, *Transportation Research*, 44E (5), 847-863.

Hudson, D., Gallardo, K. and Hanson, T. (2006) Hypothetical (non)bias in choice experiments: evidence from freshwater prawns, Working Paper. Department of Agricultural Econoimics, Mississippi State University, Mississippi State, MS, 2004.

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.

Johannesson, M., Blomquist, G., Blumenschien, K., Johansson, P., Liljas, B. and O'Connor, R. (1999) Calibrating hypothetical willingness to pay responses, *Journal of Risk and Uncertainty*, 8, 21-32.

Johansson-Stenman, O. and H. Svedsäter (2003). Self image and choice experiments: hypothetical and actual willingness to pay. Working Papers in Economics No. 94, Department of Economics, Gothenburg University.

Ladenburg, J., Olsen, S. and Nielsen, R. (2007) Reducing hypothetical bias in choice experiments, powerpoint presentation, Institute of Food and Resource Economics, University of Copenhagen, Denmark.

Landry, C.E. and J.A. List (2007). Using *ex ante* approaches to obtain credible signals for value in contingent markets: evidence from the field, *American Journal of Agricultural Economics*, 89 (2), 420-429.

List, J.A. (2001) Do explicit warnings eliminate the hypothetical bias in elicitation procedures?: evidence from field auctions for sportscards, *American Economic Review*, 91(5), 1498-1507.

List, J. and Gallet, G. (2001) What experimental protocol influence disparities between actual and hypothetical stated values? *Environmental and Resource Economics*, 20, 241-254.

List, J., Sinha, P. and Taylor, M. (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.

Available at: http://www.bepress.com/bejeap/advances/vol6/iss2/art2

Louviere, J.J., Hensher, D.A. (2001) Combining sources of preference data. In: Hensher, D.A. (Ed.), *Travel Behaviour Research: The Leading Edge*. Pergamon, Oxford, 125–144.

Lusk, J.L. (2003) Willingness to pay for golden rice. *American Journal of Agricultural Economics*, 85 (4), 840 – 856.

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

McClelland, G.H. and C.M. Judd (1993) The statistical difficulties of detecting interactions and moderator effects, *Psychological Bulletin*, 144 (2), 376-390.

Murphy, J., Allen, P., Stevens, T. and Weatherhead, D. (2004) A meta-analysis of hypothetical bias in stated preference valuation, *Department of Resource Economics, University of Massachusetts*, Amherst, January.

Murphy, J., Allen, P., Stevens, T. and Weatherhead, D. (2005) Is cheap talk effective at eliminating hypothetical bias in a provision point mechanism? *Environmental and Resource Economics*, 30 (3), 313-325.

Ohler, T., Li, A., Louviere, J. and J. Swait (2000) Attribute range effects in binary response tasks, *Marketing Letters*, 11 (3), 249-260.

Poe, G., K. Giraud and J. Loomis (2005) Simple computational methods for measuring the difference of empirical distributions: application to internal and external Scope tests in contingent valuation. *American Journal of Agricultural Economics*, 87(2), 353-365.

Portney, P.R. (1994) The contingent valuation debate: why economists should care, *Journal of Economic Perspectives*, 8(4), 3-17.

Rose, J.M. and M.C.J. Bliemer (2007) Stated preference experimental design strategies' in Hensher, D.A. and K. Button, (eds.) *Transport Modelling*, Second Edition, Handbooks in Transport, Vol. 1, Elsevier Science, Oxford, Chapter 8.

Rose, J.M., Bliemer, M.C., Hensher and Collins, A. T. (2008) Designing efficient stated choice experiments in the presence of reference alternatives, *Transportation Research* 42B (4), 395-406

Seror, V. (2007 in press) fitting observed and theoretical choices – women's choices about prenatal diagnosis of down syndrome, *Health Economics*

Sillano, M. y J. de D. Ortúzar (2005). Willingness-to-pay estimation with mixed logit models: some new evidence, *Environment and Planning* A, 37: 525-550.

Steimetz (2008) Defensive driving and the external costs of accidents and travel delays, *Transportation Research B*, in press.

Steimetz, S. and Brownstone. D. (2005) Estimating commuters' "value of time" with noisy data: a multiple imputation approach, *Transportation Research*, 39B, 865-889,

Swait, J. (2001) A Non-compensatory choice model incorporating attribute cut-offs, *Transportation Research* B, 35(10), 903-928.

Train, K. and Wilson, W. (2008) Estimation on stated-preference experiments constructed from revealed-preference Choice, *Transportation Research Part B*, 40, 191-203,

Tversky, A. and Kahneman, D. (1981) The framing of decisions and the psychology of choice, *Science*, 211, 453–458.

Wardman, M. (2001) A review of British evidence on time and service quality valuations, *Transportation Research* 37E (2–3), 107–128.