A framework for capturing heterogeneity, heteroscedasticity, non-linearity, reference dependence and design artefacts in value of time research

Stephane Hess – Andrew Daly – Thijs Dekker – Manuel Ojeda Cabral – Richard Batley

Institute for Transport Studies & Choice Modelling Centre, University of Leeds

Abstract

In early 2014, the UK Department for Transport (DfT) commissioned the first national value of travel time (VTT) study since the mid-1990s. This paper presents the methodological work undertaken for this study, with important innovations along a number of dimensions, both in terms of survey design and modelling methodology. Our findings show a rich pattern of heterogeneity across the travelling public, in terms of an impact on the VTT by both person and trip characteristics, as well as a major role for a number of characteristics that relate to the specific choices faced in a hypothetical stated choice setting, including reference dependence and non-linearities in sensitivities. We also discuss how these behavioural values were translated into values for use in appraisal, and the challenges faced when doing this on the basis of results obtained with advanced models.

Keywords: *value of time; national studies; discrete choice; random heterogeneity*

1. Introduction

Many countries conduct national value of travel time (VTT) studies to produce official values for use in appraisal (see the review by Daly et al., 2014). There have been three previous waves of such studies in Britain. First, a series of research projects during the 1960s and 1970s, the results of which were adopted and synthesised by the Department for Transport (DfT). Second, the MVA/ITS Leeds/TSU Oxford study of the 1980s leading to revised values of travel time in 1987. Third, the AHCG study of using 1994 data (AHCG, 1996) which was further analysed by ITS Leeds (Mackie et al., 2003).

Whilst routinely updated for changes in income and travel patterns, the evidence base for the values used in UK appraisal is now more than twenty years old. Along with changes in incomes, prices, demography and the mix of travel by purpose and trip length, the world has changed in other ways – the internet revolution, the quality and comfort of vehicles, working practices and, perhaps most fundamentally, the way in which people perceive time spent travelling. These substantive developments challenge the credibility of simply updating values based on old behavioural data.

Furthermore, whilst the existing set of behavioural values were obtained with methods that represented the state of the art in 1994/2003, the fields of stated choice data collection and discrete choice modelling have also seen substantial developments in the subsequent years, making the toolkit from the previous study very outdated. These developments include the growing popularity of more statistically efficient stated choice design techniques (see e.g. Rose & Bliemer, 2014, for an overview) and the availability of ever more flexible discrete choice structures (see e.g. Train, 2009, for an overview). Along with these developments in choice modelling more broadly, there have been substantial improvements to the statistical techniques used in national value of time studies, most notably starting with the work of Fosgerau et al. (2007a).

Against this background, a new study was conducted in 2014/2015[1] to meet DfT's requirement *'to provide recommended, up-to-date national average VTT measures using modern, innovative methods'*. The work conducted in the course of this study did not simply apply the best currently available methods, but made further methodological improvements. This paper presents an overview of the data collection work conducted and then focusses primarily on the modelling work undertaken, presenting our approaches to capturing reference dependence, non-linearities in preferences, and deterministic and random heterogeneity across travellers. Finally, we discuss the way in which these behavioural values were translated into values for use in appraisal, and highlight the challenges that can be faced in that context when working with advanced model structures.

The remainder of this paper is organised as follows. Section 2 gives an overview of the survey work conducted, Section 3 presents the modelling work, Section 4 outlines the results and Section 5 discusses the application of the modelling results to derive values for scheme appraisal. Finally, Section 6 presents our conclusions.

2. Survey work

Stated Preference (SP) data was collected for the three key purposes required for the appraisal framework (commute, travellers in the course of business and other non-work) and for four modes (car, rail, bus, other public transport (PT)[2]). The aim of the survey work was to provide valuations not just for travel time, but also travel time reliability and the quality of the travel experience (e.g. crowding). While work for example in Australia routinely values all these components in a single SP game, this study was required to adhere to the UK and European tradition of using a number of separate games, each looking at a subset of the journey components. Unlike previous studies, we ensured a greater representation for the more complex games, and also used joint estimation across the games.

For each purpose-mode combination, multiple SP experiments were developed, involving different unlabelled trade-offs between two alternatives described by: time/money (SP1), time/money/reliability (SP2), and time/money/quality (SP3). It should be noted that the ways in which some journey attributes, particularly reliability, are valued in the models relate to the specific appraisal framework used in the UK. Respondents received all three games, with 5 choices per game[3]. Whilst the two-game format used in previous European studies (presenting SP1 plus either

[1] The project was managed by ARUP, with ITS Leeds (in conjunction with John Bates) being responsible for the survey design, modelling work, and translation of modelled values into appraisal values. Data collection was carried out by Accent. Further information on the study as a whole can be found in the final report (Arup/ITS/Accent, 2015).

[2] Other public transport encompassed tram, light rail and underground.

[3] In order to mitigate order effects, SP1 (the easiest game) was presented initially, whilst the order of SP2 and SP3 was randomised.

SP2 or SP3) would arguably moderate cognitive burden (although those studies have consistently presented 8 choices per game), the decision to present three games was influenced by three main considerations.

Firstly, there was a desire for comparability, with values for all components being obtained from data collected from all respondents, avoiding a situation where differences in valuations across components might be due to differences in the groups of respondents supplying those valuations. Second, in order to estimate meaningful and robust values, it was judged that it would be advantageous to maximise the volume of data from games with more than two design variables (i.e. SP2 and SP3), in contrast with a two-game format that would deliver a dataset comprising around 50% SP1 observations, and 25% each of SP2 and SP3 observations. Third, we had an a priori expectation (subsequently confirmed in the results) that respondent behaviour in SP1, which is the most simplistic (and hence possibly least realistic), would be most affected by design effects, potentially reducing quality.

Where possible, with a view to enhancing realism, we 'pivoted' attribute levels around travellers' current trips, though we made some exceptions to this approach, such as for headway and crowding. The design used balance between gains and losses across the sample, as well as in terms of size of changes. The design treated time in the different levels of congestion separately, but again with symmetric pivots, ensuring that the gain-loss relationship in terms of changes in congestion was also symmetric.

2.1. SP1

SP1 used a generic format across all modes, presenting respondents with a choice between two options described only on the basis of travel time and travel cost, where one option was cheaper, but the other option was faster.

While this represents the established approach in a number of European countries (e.g. Mackie et al., 2003, Fosgerau et al., 2007a, Ramjerdi et al., 2010, Significance et al., 2013 and Börjesson and Eliasson, 2014), it is very different from the more complex approaches used in other countries (e.g. Austroad, 2006; ATC, 2006). Other than simplicity, a potential shortcoming is that the context in which time is spent is not explicitly described, and that some respondents might consider the proposition of a faster but cheaper car journey to be unrealistic, especially in a short term context in the explicit absence of tolls. Whilst we sought to contextualise the SP through the preamble instructions, it is impossible to know a priori how respondents interpret travel time in an SP1-style experiment, i.e. whether they regard it as free flow time, congested time, etc. This was a motivation for our extensive work on combining data across games and studying the differences in valuations. An important reason for maintaining the SP1 experiment was however to retain comparability with previous British studies.

2.2. SP2

SP2 also presented respondents with a binary choice, still focussing on travel cost and travel time, but presenting five different typical trip outcomes for travel time for each alternative. This is the typical approach used for reliability (i.e. unpredictable variations in travel time) in British studies (e.g. Hollander, 2006;

Batley and Ibáñez, 2013); we acknowledge that this is somewhat different from approaches used elsewhere (see e.g. Carrion and Levinson, 2012, for a review).

Informed by findings from the qualitative research, the preamble advised respondents that unreliability was associated with unpredictable (e.g. breakdowns and accidents in relation to road, or incidents on the line in relation to rail, etc.) rather than predictable variations (e.g. trips taking longer in rush hours, or fast/slow trains, etc.) in trip time. The preamble also advised that the five trips presented departed at the same time and on the same day of the week, thereby eliminating rescheduling of the trip as mitigation of unreliability.

2.3. SP3

SP3 used somewhat different presentations across modes, as follows.

- For car, the two options were described in terms of travel cost and the amount of time spent in three types of conditions (free flow, light congestion, heavy congestion).
- For rail, we presented a choice described by travel time, travel cost and the level of crowding applying to the trip. This experiment was only given to a share of respondents, with the remainder being given an operator choice game not covered by this paper.
- For bus, two different experiments were also used. For the first group, we presented a crowding game analogous to the rail game, albeit with different crowding definitions, while the second group received a choice between two bus routes described in terms of free flow time, slowed down time, dwell time, headway and fare.
- For other PT, we presented a crowding game analogous to the bus game. This experiment was only given to a share of respondents, with the remainder being given a mode choice game not described in this paper.

2.4. Experimental design

The field of experimental design has seen substantial developments over recent years (see, e.g., Rose and Bliemer, 2014), with efficient designs using prior information on parameter values leading to more meaningful trade-offs that increase the information content in the data. For our study, we used priors based on an extensive review of values obtained in past studies, especially the detailed meta-analysis work on British values of travel time of Abrantes and Wardman (2011).

The designs produced for this study follow the state of the art in the field, making use of D-efficient designs, where we relied on designs for MNL models. While this may lead to lower efficiency than an approach optimising the designs for data with random heterogeneity, we felt this was an acceptable simplification in the absence of reliable priors for heterogeneity. We however allowed for uncertainty in our priors by using Bayesian D-efficient designs. We worked with wide regions, using normally distributed priors, with standard deviations that were 50% of the mean values. We also avoided the inclusion of strictly dominant alternatives by using a regret measure, see Bliemer et al. (2014).

Different designs were produced across purposes and for different journey lengths (i.e. different reference values). While it is clear that different designs are needed for different games (e.g. separate design for time-cost and for time-cost-reliability trade-offs), it is important to recognise that an efficient design is optimised for the specific values of attributes and priors used in the design. This had two separate dimensions in the present context, with different designs by trip purpose and for a set of different representative trips. Each respondent was then given a design based on the trip closest to their reference trip, with percentage variations (or pivots) applied to the specific reference trip for that person; the pivots were obtained from the design. The number of reference trips used varied by purpose, with the lowest number for bus (2) and the highest number for rail (20). In total 315 designs were produced for this study.

2.5. Sampling, recruitment and data cleaning

The study as a whole involved a number of surveys, although this paper focuses upon the general public SP survey, which was by far the most substantial survey conducted. This survey was recruited principally through the interception of travellers (80%), complemented by some telephone recruitment (20%), where this mixed approach was designed to capture a range of journey distances. The intercept survey was administered by Computer Aided Personal Interview (CAPI) using Android tablets. Interviewers approached a random sample of travelling adults (typically 1 in 3) and asked scoping questions to check whether the respondent was in scope and matched required quotas. For the general public telephone sample, random digit dialling (RDD) was used to give a geographically representative sample of the population of England as shown in the 2011 Census by region. Irrespective of the recruitment method, respondents were able to complete the survey on-line or by means of an operator-led telephone call. Intercept-based recruitment achieved an overall response rate of 37%, whilst telephone recruitment achieved a response rate of 61% from those recruited through RDD.

In total, responses from a sample of 8,623 different individuals were collected, split across the three purposes and four modes, with car and rail having the largest samples. The data from the field surveys was subjected to detailed examination in order to establish its quality and reliability for modelling. Our overall approach to data cleaning was that records were removed only when absolutely necessary and with a view to avoiding bias. In general, the most common cleaning factors were missing distance information and missing cost information, as well as exclusions driven by unrealistic cost information, which would have made the SP choice scenarios for those respondent unrealistic. A final sample of 7,692 respondents was retained for analysis.

A number of basic diagnostic tests were performed relating to a number of behavioural traits that can adversely affect model estimation (Hess et al., 2010). Our observations for these measures are as follows:

- Rate of left or right non-traders: the proportion of people who always chose the option presented on the same side across all SP choice tasks was found to be negligible.
- Rate of time non-traders: in most sub-samples, only 1-2% of the respondents consistently chose the fastest travel option. This gives reassurance that the ranges presented in the survey were wide enough, and suggests that we can be somewhat less concerned about the ability to estimate the tail of the VTT distribution in Mixed Logit models (Börjesson et al., 2012).

- Rate of cost non-traders: in most sub-samples, only 5% of people consistently chose the cheapest option across all tasks, which is lower than in many other studies and supports the ranges presented in the trade-offs (cf. Hess et al., 2010).

3. Modelling approach

The modelling approach used in our study has a number of distinct methodological components that we will now look at in turn. Each time, where required, we look separately at the treatment required in the different types of SP games. The individual methodological components address different behavioural phenomena that are potentially at play at the same time, requiring a complex modelling framework. We first look at the error structure in the model before discussing the treatment of size and sign effects and the use of a joint estimation approach across games. We finally turn to the incorporation of deterministic and random heterogeneity and discuss how this needs to be implemented in the case of joint estimation across games with different survey contexts, error structures and size and sign effects.

3.1. Multiplicative vs additive error structures

The utility in a choice model is decomposed into a deterministic and a random component, the error term. Typically, following Daly and Zachary (1975) who pioneered this formulation for estimating valuations, models make use of standard additive error structures, where $U = V + \varepsilon$, with V and ε giving the deterministic and random components of utility, respectively. This means that the error term is 'white noise' and the amount of error is independent of the deterministic utility and hence of the characteristics of the choice that is modelled. While this is widely used in practice, it is an assumption that may not be valid in many circumstances.

In the present study, we move away from this assumption by relying on models using a multiplicative formulation (Harris and Tanner, 1974; Fosgerau and Bierlaire, 2009). In a general multiplicative formulation, we replace the typical additive specification of the utility of an alternative by $U=V.\varepsilon$, where V and ε are still defined as the deterministic and random components of utility, respectively. That is, the random (error) component of utility is taken to multiply the deterministic component, rather than be added to it. However, to operationalise the multiplicative model, we follow Fosgerau and Bierlaire (2009) in reformulating the model through a logarithmic transformation, i.e. $\log(U)=\log(V)+\log(\varepsilon)$, which is possible because the log transform is strictly monotonic. To implement this formulation we need to ensure that both V and ε are positive. Issues concerning V are discussed in the context of specific models; we guarantee that ε is positive by assuming that it follows a log-extreme-value distribution, again following Fosgerau and Bierlaire (2009). The practical advantage given by the multiplicative approach is that it becomes much easier to make an assumption of approximately constant variance for ε. In general, it is found that utility variance increases as utility increases and this is handled automatically in the multiplicative form of the model.

The multiplicative formulation represents the state of the art in VTT estimation for experiments of the SP1 type. A corresponding approach for SP2 and SP3 is also possible, as used in the recent Danish national VTT study (Fosgerau et al., 2007a), but further developed here for reference dependence. Extensive empirical testing showed the multiplicative specification to be superior to the additive specification for our work, across all games.

For SP1, the analysis is quite simple because there are just two attributes: time and cost. For this reason, it is possible to formulate the econometric model in a multiplicative log willingness-to-pay form, which has proved successful in several previous studies (Fosgerau et al., 2007b; Ojeda-Cabral et al., 2016). This involves working with the logs of time and cost differences, rather than working with the logs of two utility functions each given by the sum of a contribution made by time and cost for that alternative.

Specifically, we now use (avoiding for now additional subscripts for respondents and choice tasks):

$$V_s = \mu_{SP1}.\log\left(-\frac{cost_1 - cost_2}{time_1 - time_2}\right) \qquad (1)$$

$$V_e = \mu_{SP1}.\log\omega \qquad (2)$$

where the alternatives are re-ordered so that V_s gives the utility of the cheaper but slower option and V_e gives the utility of the faster but more expensive option, ω is the estimated VTT (for reductions in time, i.e. a positive value of time) and μ_{SP1} is an estimated scale parameter. If the VTT is greater than the implied boundary VTT (given by the cost difference divided by the time difference) presented to the respondent, then the faster option will be chosen, otherwise the cheaper option is chosen.

With this specification, the error in the models is proportional to the boundary value of time, i.e. the trade-offs faced by respondents. This is consistent with the notion that the main source of error in simple time-money trade-offs would be unexplained heterogeneity in VTT measures, which would thus lead to larger error in scenarios where the value of time required to choose the expensive option is larger. This specification was shown to not only outperform the simple additive structure, but also a structure where the error is proportional to overall utility (cf. later discussions for SP2 and SP3).

We now have that the probability of the observed sequence of T ($t = 1, \ldots T$) choices for person n for SP1 is given by the product of logit probabilities:

$$P_{SP1,n} = \prod_{t=1}^{T}\left(\frac{e^{V_{snt}}}{e^{V_{snt}} + e^{V_{ent}}}\right)^{\delta_{snt,SP1}}\left(\frac{e^{V_{ent}}}{e^{V_{snt}} + e^{V_{ent}}}\right)^{\delta_{ent,SP1}} \qquad (3)$$

where $\delta_{snt,SP1}$ is 1 if and only if the slower/cheaper option is chosen by respondent n in task t, with a corresponding definition $\delta_{ent,SP1}$ applying to the faster/expensive option.

It is important to note that, with a Multinomial Logit (MNL) specification of the model, the use of ω alone as an estimate of the VTT is likely to underestimate the true mean VTT. Indeed, in the model above, the error term is likely to capture not just noise but also heterogeneity in the VTT (given that this model works in relative valuations)[4].

[4] Using the extreme example that all the error term relates to heterogeneity in the value of time, we would have that $VTT = \omega e^{\frac{\varepsilon_1 - \varepsilon_2}{\mu_{SP1}}}$, where ε_1 and ε_2 are the log-extreme value errors. As such, for an MNL model, ω relates more to a median than a mean VTT.

Given the reliance on Mixed Logit models later in the analysis, where we explicitly incorporate random heterogeneity in VTT, this issue largely disappears, as the additional random components distributed across respondents then capture random variation in VTT.

For SP2 and SP3, log willingness-to-pay is not a feasible approach as there are multiple attributes and the signs of differences from the reference value are not consistent. In this case, we make the error proportional to the overall utility, which essentially means the model implies greater error on longer trips.

In an additive model, we would have that:

$$V_j = \tau_{cost} cost_j + \sum_k \tau_k x_{jk}, \quad (4)$$

where x_{jk} refers to K different non-cost attributes for alternative j, and where the τ parameters are estimated marginal utilities. The τ parameters would be negative for undesirable attributes, and positive for desirable attributes. This can be rewritten for a mathematically equivalent specification (Train and Weeks, 2006) in valuation space as:

$$V_j = -\mu_{SPx}\left(cost_j + \sum_k \omega_k x_{jk}\right), \quad (5)$$

where μ_{SPx} is a positive scale parameter for SPx (x=*2*,*3*), and where ω_k is now a directly estimated monetary valuation for changes in x_{jk}[5]. The negative sign on the entire utility means that the ω_k are positive for undesirable attributes, i.e. they relate to a willingness-to-pay for avoiding positive changes in an attribute. This is appropriate for 'bads' such as time but not for 'goods'. It then makes sense to replace $\sum_k \omega_k x_{jk}$ by $\sum_{k_T} \omega_{k_T} x_{jk_T} - \sum_{k_{NT}} \omega_{k_{NT}} x_{jk_{NT}}$, where x_{jk_T} are bads and $x_{jk_{NT}}$ are goods. For our analysis, we include attributes such as travel time variability, headway and delays in x_{jk_T}. In the multiplicative model, we would use:

$$V_j = -\mu_{SPx} \cdot \log\left(cost_j + \sum_{k_T} \omega_{k_T} x_{jk_T} - \sum_{k_{NT}} \omega_{k_{NT}} x_{jk_{NT}}\right) \quad (6)$$

where ω_{k_T} remains the directly estimated monetary value (i.e. willingness-to-pay) for reductions in bads, and $\omega_{k_{NT}}$ is the directly estimated monetary value for increases in goods $x_{k_{NT}}$ (i.e. a priori, both parameters would be expected to be positive).

Again using a log-extreme value distribution for ε, the probability for the observed sequence of choices for respondent n in SPx is now given by:

$$P_{SPx,n} = \prod_{t=1}^{T} \prod_{j=1}^{2} \left(\frac{e^{V_{jnt}}}{e^{V_{1nt}} + e^{V_{2nt}}}\right)^{\delta_{jnt,SPx}} \quad (7)$$

where $\delta_{jnt,SPx}$=1 if and only if alternative j is chosen by respondent n in task t in SPx.

For SP2, an additional complication arises as the respondents are presented, for each alternative, with five different possible outcomes in terms of travel time. For car, we simply work with the mean travel time and the standard deviation across

[5] Of course $\mu_{SP2} = -\tau_{cost}$ and $\omega_k = \tau_k / \tau_{cost}$.

outcomes. However, for public transport, we work with the "usual travel time" (which is effectively the mean of the five outcomes) as well as valuations for early and for late arrivals (unscheduled, as opposed to schedule delay). This approach was driven by the conventions of official transport demand forecasting and appraisal guidance in the UK (e.g. DfT, 2013; ATOC, 2012), and was not intended to exploit recent academic developments in estimating the value of reliability (e.g. Tseng and Verhoef, 2008; Fosgerau and Karlström, 2009; Engelson and Fosgerau, 2016).

We would thus get that the value for outcome 1 (out of 5) for alternative j, say $v_{j,1}$ is given by:

$$v_{j,1} = cost_j + \omega_{time} time_j + \omega_{early} earliness_{j,1} + \omega_{late} lateness_{j,1} - \omega_{late\ penalty} late_{j,1} \tag{8}$$

where $time_j$ is the usual travel time, $earliness_{j,1}$ and $lateness_{j,1}$ relate to the amount of early or late delay in outcome 1 for alternative j, and $late_{j,1}$ is equal to 1 if the first outcome has late arrival, and 0 otherwise. The estimates for ω_{time}, ω_{early} and ω_{late} are three distinct value of time measures. It should be noted that ω_{early} and ω_{late} relate to unexpected early or late arrival, i.e. not schedule delay (a planned arrival which is earlier or later than the preferred arrival time for that traveller). As such, we would expect a positive value for ω_{late} (just as for ω_{time}), as reductions would imply shorter travel time, while $\omega_{late\ penalty}$ is the willingness-to-pay for late arrival, which we expect to be negative. For ω_{early}, we also expect a negative estimate as reductions in earliness would mean longer travel time, i.e. respondents should desire an increase in earliness. There should be no strong a priori expectation that the estimates of ω_{early} and ω_{time} are simple opposites, given non-linearities in sensitivities, but also given potentially different behavioural response to scheduled travel time and an unscheduled early arrival.

On the assumption of equal weight being given to the five outcomes, we can then use:

$$V_j = -\mu_{SP2} \cdot \log\left(\sum_s \frac{v_{j,s}}{5}\right) \tag{9}$$

where s is an index over the five possible outcomes. Building on the work of Liu and Polak (2007), we allow for differential weights for the five outcomes using the constant absolute risk aversion (CARA) specification, with:

$$V_j = -\mu_{SP2} \cdot \log\left(\sum_s \frac{1-e^{-\alpha v_{j,1}}}{\alpha} \frac{1}{5}\right) \tag{10}$$

where, with α approaching zero, we get a risk neutral model, positive α implies risk averseness, with the opposite applying for negative α.

A similar specification for car was not possible in the context of an appraisal framework requiring a standard deviation of travel time (which obviously does not apply at the individual outcome level).

3.2. Treatment of size and sign effects

Many SP-based VTT studies have found that the values obtained depend on the sign (i.e. asymmetries) and size (i.e. non-linearities) of time and cost changes *relative* to a 'reference' value (e.g. Börjesson and Eliasson, 2014; De Borger and Fosgerau, 2008 (dBF)). These findings can be related to Prospect Theory, e.g. that gains are attributed a lower absolute value than equivalent losses (Kahneman and Tversky, 1979).

We have adopted the principles of the approach to modelling reference dependence set out by dBF, and have further developed it for our study. In a model, we wish to introduce the concept of gains and losses, as well as the basic difference between the alternatives. The question arises as to whether we want to measure differences from the base value or differences between the alternatives.

Considering the degrees of freedom, for a given choice task, there are just three measurements of attribute values: the current (reference) value and the values presented for the two alternatives. Any attempt to introduce more than three variables by calculating differences etc. is liable to fail by introducing overspecification.

In previous studies such as the 1994 British data and the more recent Scandinavian studies, these effects (differences between alternatives and differences against the reference trip) would have been perfectly confounded as the reference values for both time and cost appeared in every choice task. Previous studies have interpreted the estimates of the size effects as relating to the differences between the alternatives, but the alternative interpretation would have been equally plausible. Given the vast evidence on the importance of reference dependence, our work focusses on that, rather than on differences between the alternatives, an approach also supported by empirical tests on the data.

The concept is to introduce non-linear functions that express the possibility that size and sign effects exist. This is done by defining a function that gives the value of a change Δx relative to the reference value x_0 of a given attribute, where, following dBF:

$$v(\Delta x) = S(\Delta x).\exp(\eta\, S(\Delta x)).|\Delta x|^{\alpha} \tag{11}$$

with $\Delta x = x - x_0$, $\alpha = 1 - \beta - \gamma S(\Delta x)$

$S(\Delta x)$ is the sign function, defined for $\Delta x \neq 0$ by $S(\Delta x) = \Delta x\ /\ (|\Delta x|)$, i.e. it takes the values ± 1 with the same sign as Δx; for convenience we also specify that $S(0) = 0$.

η gives the difference of gain value and loss value from an 'underlying' value. It is explicitly assumed by dBF that gains and losses exactly bracket this underlying value. The parameter η measures the sign effect. It is expected that $\eta > 0$, so that the value of losses (increases in Δx) is greater than the value of gains.

β allows the impact of gains and losses to be non-linear. If $\beta > 0$, the marginal value of changes decreases as the change increases, i.e. the value is ‘damped’. This is the main measure of the size effect. Generally, we anticipate that β should be larger for cost than for time, so that VTT increases as the changes increase, while small time savings have lower monetary value.

γ allows the non-linearity of value to be different for gains and losses. Essentially, this gives an interaction between the sign and size effects. A negative value for γ would for example mean that any damping (i.e. decreasing marginal effects for larger changes) would be smaller for increases (losses) than for decreases (gains) from the reference value, or that any increasing sensitivity for larger changes (i.e. with $\alpha > 1$), would be stronger for increases than for decreases from the reference value.

The value functions are defined to have arguments and values denominated in cost units. Thus the cost value of a cost change Δc is given by $v(\Delta c)$, while the cost value of a time change t is given by $v(\theta\Delta t)$, where θ is the ‘underlying’ value of time. Differently from the dBF work, and from how it was used in the Danish work, we are also able to estimate separate η, β and γ parameters for both time and cost in SP1 as our design does not impose the presence of the reference value for one of the two alternatives in the choice task[6].

We next turn to how the actual VTT can be calculated within the above framework. A simple way to see the derivation of VTT (and other WTP measures) is to think of the values of Δc and Δt that would maintain indifference with the base situation in which $\Delta t = \Delta c = 0$ and the total value is of course zero. Thus when we have a specific value $\Delta t'$, and we have estimated the parameters of the value functions v, we can find the value $\Delta c'$ such that $v(\Delta c') + v(\theta\Delta t') = 0$. The average willingness-to-pay per unit of time is then $\Delta c'/\Delta t'$.

It is reasonable to extend the method of dBF in taking the average of the gain value and the loss value to express an ‘underlying’ VTT. In fact, it is difficult to formulate an alternative: in the SP context we obtain gain values and loss values and these need to be averaged in some way, in part also as we do not know whether these effects may be amplified in an SP setting (compared to real life). That is, to obtain a reference-free value we need to calculate the average of the loss value of a given Δx and the gain value of the same Δx to obtain a reference-free value of Δx. As in dBF, we calculate the *geometric* mean[7] of $v(\Delta x)$ and $-v(-\Delta x)$:

$$\sqrt{v(\Delta x).-v(-\Delta x)} = \sqrt{\exp(\eta\,).|\Delta x|^{1-\beta-\gamma}.\exp(-\eta).|\Delta x|^{1-\beta+\gamma}}$$

$$= \sqrt{|\Delta x|^{2-2\beta}} = |\Delta x|^{1-\beta} \qquad (12)$$

[6] The DATIV data used by Fosgerau et al. (2007a), like previous British and several other European studies, was based on a design in which the current time and cost always appeared in one or other of the alternatives presented. This design has the effect that it is not possible to make separate identifications of β for both time and cost. However, with the new British data the design is more varied and it is possible to make these identifications. This was also the case in work conducted by Hjorth & Fosgerau (2012) on the Norwegian data.

[7] dBF assumed (unnecessarily) that γ was zero to calculate the geometric mean.

Thus, finding the geometric mean of the gain and loss values leads to an estimate of a 'reference free' value in which η and γ do not appear. However, there is no analogous argument to eliminate β and the value remains a function of Δx.

Solving the equation $v(\Delta c') + v(\theta\Delta t') = 0$, for the gain-loss average value functions $v(\Delta x) = S(\Delta x)|\Delta x|^{1-\beta}$ we obtain, for oppositely signed Δc and Δt,

$$|\Delta c|^{1-\beta_c} = (\theta|\Delta t|)^{1-\beta_t} \tag{13}$$

$$|\Delta c| = (\theta|\Delta t|)^{\frac{1-\beta_t}{1-\beta_c}} = (\theta|\Delta t|)^{\kappa} \tag{14}$$

where $\kappa = \frac{1-\beta_t}{1-\beta_c}$, so we can calculate the VTT (per unit of time) as:

$$VTT = {|\Delta c|}/{|\Delta t|} = \theta^{\kappa}|\Delta t|^{\kappa-1} \tag{15}$$

Here it is obvious that if $\beta_c = \beta_t$, VTT is independent of Δt, as the time and cost damping cancel out, i.e. we get that $\kappa = 1$. However, in general the β values will not be equal and VTT is not equal to θ. It is for this reason that we change the notation from ω in the non-reference-dependent models, which is always the VTT, to θ in these models, noting that the estimate of θ then needs to be used in (15) alongside that for κ to calculate the VTT.

In the formulation using value functions it is not appropriate to obtain VTT for finite time differences from strictly marginal valuations, as would be found by differentiation. The concept is to determine the value of a finite amount of time Δt, where the marginal value of both time and cost varies continuously. The use of differentials would, for instance, give the value of changing Δt from 10 to 11 minutes, whereas what is required is the value of the change from 0 to 10 minutes. Moreover, the differential of the time value depends only on Δt, whereas the differential of the cost value depends on both Δt and Δc, so that the ratio of differentials varies in two dimensions.

Using the Δ notation and subtracting $(\mu_{SP1} \log \omega)$, (1) and (2) can be reformulated without changing their meaning as

$$V_1 = \mu_{SP1}.\log\left(-\frac{\Delta c_1 - \Delta c_2}{\omega\Delta t_1 - \omega\Delta t_2}\right) \tag{16}$$

$$V_2 = 0 \tag{17}$$

This purely technical reformulation allows us to extend the model to include reference dependence.

For SP1 the design ensures that the value differences have opposite signs themselves. Comparing alternatives s and e, respondents value the cost difference they are offered by $\left(v(\Delta c_1) - v(\Delta c_2)\right)$ and the time difference by $\left(v(\theta\Delta t_1) - v(\theta\Delta t_2)\right)$. It is then 'rational' to choose the slow alternative if $|v(\Delta c_1) - v(\Delta c_2)| > |v(\theta\Delta t_1) - v(\theta\Delta t_2)|$. This implies a model form:

$$V_s = \mu_{SP1}.\log\left(-\frac{v(\Delta c_1) - v(\Delta c_2)}{v(\theta\Delta t_1) - v(\theta\Delta t_2)}\right) \tag{18}$$

$$V_e = 0 \tag{19}$$

This is the model that is estimated, with separate dBF parameters for time and for cost, using the reported time and cost for the respondent's recent trip as the reference points. It is of course possible that respondents use a different reference point (see e.g. the work of Hess et al., 2012), but the survey did introduce the choice tasks by asking respondents to think of their recent journey, and the use of these reference points is in line with most other empirical applications in the field. The effects we retrieved were meaningful and seem to support our approach.

To introduce reference dependence in the specifications for SP2 and SP3, we can replace any of the terms in the utility functions by the corresponding v for changes in the associated attribute, noting that if η, β, γ are constrained to zero, we get $v(\theta\Delta x) = \theta\Delta x.$ For the example of a case with cost, time and delay (d), we could write:

$$V_j = -\mu_{SP\mathrm{x}} \log\left(c_j + \theta_t t_j + \theta_d d_j\right) \tag{20}$$

and we could substitute value functions for these components:

$$V_j = -\mu_{SP\mathrm{x}} \log\left(v\left(\Delta c_j\right) + v\left(\theta_t \Delta t_j\right) + \ v\left(\theta_d \Delta d_j\right) + \theta_t t_0 + \theta_d d_0 + c_0\right) \tag{21}$$

The inclusion of the base values in addition to the value functions inside V_j is required, as, in contrast with the model for SP1, we are not working in relative valuation space.

We can obtain the required generalisation by estimating or eliminating some or all of the parameters expressing reference dependence. In each case θ relates to the willingness-to-pay for changes in the specific attribute.

An important discussion relates to the choice of reference values for the individual non-cost attributes[8] in SP2 and SP3, where the situation is not as straightforward as for SP1.

- For car SP2, we allowed for reference dependence only for travel time, in the absence of a value for travel time reliability for the reference trip.
- For the three public transport SP2 games, we used the time for the reference trip as the reference value for the *usual* travel time for each of the five outcomes, with no reference points being available for early or late arrival.
- For car SP3, we initially attempted the use of separate reference values for the three individual time components, but did not obtain conclusive results in relation to reference dependence, possibly due to respondents finding it difficult to report an accurate breakdown of congestion across parts of their journey. Better fit and more reasonable results were obtained by applying reference dependence to the total travel time, thus summing up the three components, while allowing for different θ values for the three valuations within the total travel time. For a given alternative *j* in task *t*, the value function for total time would then use:

$$\theta_{TT,j} = \frac{\theta_{FFT} FFT_{jt} + \theta_{LCT} LCT_{jt} + \theta_{HCT} HCT_{jt}}{FFT_{jt} + LCT_{jt} + HCT_{jt}} \tag{22}$$

[8] Cost was obviously treated the same way as in SP1.

- For SP3 crowding games, we used the travel time for the reference trip as the reference value, with the specific θ being used depending on the crowding level presented, i.e.

$$\theta_{TT,j} = \sum_{k=1}^{K} \theta_{T,crowding_k} \delta_{crowding_k,j} \quad (23)$$

where $\delta_{crowding_k,j} = 1$ if and only if crowding level *k* applies for alternative j.

- For the bus time components SP3 game, no reference dependence was used for headway, while, for the three travel time components (free flow, slowed down and dwell time), an approach corresponding to that for car SP3 above was used.

3.3. Joint modelling approach

We allow for differences in valuations across games by using separate multipliers for each valuation in our models, relating them to the base θ, say θ_0. This also tests for differences in interpretation for attributes that are common across games. Using the example of car, we then obtain six separate θ measures, e.g. $\theta_{SP1,VTT} = \zeta_{SP1,VTT}\theta_0$, for the valuation of travel time in SP1, and $\theta_{SP2,VTT} = \zeta_{SP2,VTT}\theta_0$, for the valuation of average travel time in SP2, with the others being the standard deviation and the three separate travel conditions. A normalisation is required here, and we therefore set $\zeta_{SP1,VTT} = 1$, meaning that the base valuations relate most directly to SP1.

We also make use of game-specific error scale parameters, as already outlined in the utility specifications for the separate games. With $P_{SP1,n}$, $P_{SP2,n}$ and $P_{SP3,n}$ being the likelihood of the observed set of choices in the three sets of stated choice scenarios, the joint probability of the choices observed for respondent *n* is given by:

$$P_n = P_{SP1,n} P_{SP2,n} P_{SP3,n} \quad (24)$$

The main benefit of joint estimation across games is increased robustness for those parameters shared across games, which in our case is the set of covariates explaining deterministic heterogeneity in valuations as well as the random heterogeneity parameters. But it is also true that if parameters are not in fact the same across datasets, then there is no robustness to gain. In the context of a purely academic study, one would test statistically whether it is appropriate to pool datasets. However, in the present context, there was an a priori requirement to ensure consistency of values across games, as it would not be acceptable for implementation to have different income elasticities across different value of time components, for example. We do, as mentioned above, allow for difference in the base valuations across games, and enforce equality only in the covariates and in the degree of variation (in terms of the relationship between mean and variance) of the individual value of time measures. In Section 3.6, we provide some limited empirical evidence to offer support for this specification.

3.4. Inclusion of deterministic heterogeneity

An extensive specification search was undertaken to test the impact on valuations of a substantial range of person and trip covariates, as well as to account for potential design effects. Alongside size and sign effects, this latter group included testing for impacts of the position of the time attribute relative to the cost attribute, and the impact of whether the cheaper option was shown on the left or right.

All the person and trip characteristic effects were interacted in the same way for the individual θ measures used in different games. This equates to an assumption that their impact is consistent across the different types of components valued in our work, and that they relate primarily to an underlying willingness-to-pay, independently of the good being valued. This is of course a simplification, but one that was necessary in the context of this project.

For the majority of the components above, multipliers on the VTT were estimated, with one category for the attribute being used as the base, for which the multiplier was then set to a value of 1. Using gender as the simplest example, we would then for example multiply θ by $(\zeta_{female} female_n + male_n)$, where ζ_{female} is an estimated multiplier on the VTT for female respondents (i.e. $female_n = 1$ if respondent *n* is female), using male as the base.

A different specification was used for four continuous effects, namely income and the cost, time and distance of the reference alternative, where an elasticity specification was used. Taking income as the example, the multiplier on θ would be given by:

$$\left(\frac{inc}{40}\right)^{\lambda_{inc}} \delta_{income\ reported} + \zeta_{not\ stated}\delta_{income\ not\ stated} + \zeta_{unknown}\delta_{income\ unknown} + \zeta_{refused}\delta_{refused} \tag{25}$$

With this specification, inc is a continuous income variable (expressed in thousands of pounds per annum), λ_{inc} is an estimated income elasticity, $\zeta_{not\ stated}$, $\zeta_{unknown}$ and $\zeta_{refused}$ are multipliers on the VTT for respondents with unreported income, and the four δ terms are dummy variables (one of which is set to 1) categorising the respondents according to whether income was reported or not. The value of 40 chosen as a denominator simply means that the base θ relates to a respondent with an annual income of £40,000. Tests were conducted to determine which of the different income variables was most appropriate for given purposes, where, across modes, we ended up with a specification using household income for commuting and for other non-work, with personal income used for business.

A corresponding specification was used to estimate elasticities of θ with respect to the cost of the reference trip (with a base of £5), the travel time of the reference trip (with a base of 30 minutes) and the distance of the reference trip (with a base of 20 miles).

Given that any impacts of the positioning of the time/cost attributes and the cheap/expensive alternative are purely SP effects which we do not want to influence the estimated VTT, a multiplicative effects coding approach was used. Further, these effects were entered at the level of individual choices in individual games, unlike the other covariates. Taking the example of whether the cheap option was presented on the left or the right, the additional multiplier on the value of the game and task specific θ in choice task *t* would be given by:

$$\zeta_{cheap\ left}\delta_{cheap\ left_t} + \frac{1}{\zeta_{cheap\ left}}\left(1 - \delta_{cheap\ left_t}\right) \tag{26}$$

where $\delta_{cheap\ left_t}$ is set to 1 if and only if the cheap option is presented on the left in task t. This specification ensures that the base estimates of θ relates to the average situation (geometric mean) in the data according to how often the cheap option is presented on the left or on the right.

Other than the interaction of the covariates with θ, we also allowed for the order of the games to be interacted with the scale parameters.

Finally, we also tested the inclusion of constants for the alternative presented on the left directly in the utility functions in SP2 and SP3, along with a constant for any alternatives with no travel time variability in the SP2 games. These terms, by being entered directly into the utility functions, do not affect the VTT measures.

The inclusion of generic covariate effects across different valuations is complicated by the role of the β parameters (remembering that η and γ do not enter into the WTP calculations). Using the VTT in car SP1 as an example, we would have:

$$\theta_{SP1,VTT} = \theta_0 \left(\zeta_{SP1,VTT} \prod_m z_m^{\lambda_m} \prod_n \zeta_n^{z_n}\right)^{1/\kappa_{SP1,VTT}} \tag{27}$$

where θ_0 would relate to a respondent with the base values for all covariates in a single game model, λ_m is the elasticity for a continuous covariate z_m, ζ_n is the multiplier applied for a discrete covariate z_n when its value is 1, and m and n run over the continuous and discrete covariates respectively.

The inclusion of the exponent $1/\kappa_{SP1,VTT}$ ensures that the impacts λ and ζ apply directly to the VTT for SP1, as:

$$\begin{aligned} VTT_{SP1} &= \theta_{SP1,VTT}{}^{\kappa_{SP1,VTT}} |\Delta t|^{\kappa_{SP1,VTT}-1} \\ &= \theta_0^{\kappa_{SP1,VTT}} \zeta_{SP1,VTT} \prod_m z_m^{\lambda_m} \prod_n \zeta_n^{z_n} . |\Delta t|^{\kappa_{SP1,VTT}-1}, \end{aligned} \tag{28}$$

This formulation ensures that the estimates of the impacts of the covariates as well as the game-specific multipliers relate directly to all the individual valuations, as do the standard errors. This re-parameterisation is required given that the value of κ can differ across games and across valuations within a given game.

As a result of the differences in size and sign effects across valuations, the game-specific multipliers, e.g. $\zeta_{SP1,VTT}$, can also not be directly understood to explain the differences in the valuations across games. Indeed, the differences in say the VTT between SP1 and SP2 would be given by:

$$VTT_{SP1}/VTT_{SP2} = \zeta_{SP1,VTT}/\zeta_{SP2,VTT}\ \theta_0^{\kappa_{SP1,VTT}-\kappa_{SP2,VTT}} |\Delta t|^{\kappa_{SP1,VTT}-\kappa_{SP2,VTT}} \tag{29}$$

which is thus not as simple as $\zeta_{SP1,VTT}/\zeta_{SP2,VTT}$. While $\theta_0^{\kappa_{SP1,VTT}-\kappa_{SP2,VTT}}$ is simply a constant which can be calculated, $|\Delta t|^{\kappa_{SP1,VTT}-\kappa_{SP2,VTT}}$ is a function of Δt, and as a result, the ratios of different valuations depend on the assumptions made in relation to Δt. This point is addressed in Section 5 to follow.

3.5. Incorporating random heterogeneity

We finally allow for random heterogeneity in θ_0, i.e. the base value before the incorporation of covariates and reference dependence. This means that the contribution to the likelihood function by person *n* is now given by:

$$P_n = \int_{\theta_0} P_{SP1,n}(\theta_0) P_{SP2,n}(\theta_0) P_{SP3,n}(\theta_0)\, f(\theta_0) d\theta_0 \tag{30}$$

where $f(\theta_0)$ is the density function for θ_0.

For the present study, after extensive testing, we settled on the use of a log-uniform distribution, which has a shorter tail than the lognormal distribution. While used in tests by Fosgerau (2006), our work seems to present the first large scale application using this distribution. In the same way that a variate *x* has a lognormal distribution if $y = log(x)$ is normally distributed, we define *x* as log-uniformly distributed if $y = log(x)$ is uniformly distributed.

Denote *a* as the lower bound and *b* as the spread of a uniform distribution, then the mean of the resulting log-uniform distribution is given by:

$$E(\theta_0) = \frac{\exp(a+b) - \exp(a)}{b} \tag{31}$$

and the variance by:

$$Var(\theta_0) = \exp(2a)\left[\frac{\exp(2b)-1}{2b} - \frac{(\exp(b)-1)^2}{b^2}\right] \tag{32}$$

With the above specification, the heterogeneity is entered at the respondent level rather than the game level, meaning that the integration over the distribution of θ_0 is carried out over all choices for the respondent in the joint games. We estimated the model using simulated log-likelihood, with 500 Halton draws used per respondent. To investigate the appropriateness of using the log-uniform distribution, we conducted tests showing improvements in fit over the lognormal distribution, as well as a lack of substantial improvements by adding semi-nonparametric terms using the Fosgerau & Mabit (2013) approach (detailed results available on request).

The Danish and Swedish studies discussed the issue of tails of random distributions in great detail. They recognised that while allowing for random heterogeneity yields important benefits, it can also lead to issues with the right hand tail of the value of time distribution, where the tail obtained by the parametric distribution may have little or no support in the data. In particular, they observed a non-trivial share of the estimated distribution being substantially higher than the highest presented trade-offs. While this problem can be in part addressed by presenting wider trade-offs (as in the Swedish study), they also showed that a need may arise to truncate the estimated distribution. In the present study, we proceeded without truncation or censoring of results for a number of key reasons.

Firstly, censoring is an inherently unsatisfactory process. It implies the estimation of a model and then changing the outputs from that estimation process with a view to rejecting a few ‘inconvenient’ values. It is then possible that the results for other model parameters do not relate to the censored results; i.e. the estimated model no longer gives an optimal fit to the data. In other words, if say censoring was applied during estimation for the value of time distribution, then it is also likely that different values would be obtained for key covariates such as income elasticities.

Secondly, our models were estimated jointly across all games, while the other studies used random heterogeneity only at the level of an individual game, in particular SP1, which would substantially increase the scope for non-trading on time and lead to long tails of the estimated distribution. The joint estimation on all games also prevents us from easily conducting non-parametric analysis of the data as performed in the Scandinavian work (Fosgerau, 2007), i.e. inferring the distribution of the VTT from ‘looking’ at the data.

Thirdly, the use of the log-uniform distribution itself reduces the issues with extreme tails, with the upper limit not being infinity.

Fourthly, the calculation of an appropriate censoring point would be arbitrary and, given the joint estimation across three games, could not use the simple boundary VTT from SP1, as was done in the Scandinavian work.

Finally, the wide ranges used in our design work gave extensive coverage to the domain of possible VTT values to be revealed in the models.

To further investigate the suitability of the ranges offered, we conducted the test proposed by Fosgerau et al. (2007a) in the Danish study. This test uses the final estimated model and applies it to the choice data for SP1 to compute the probability for the expensive alternative in each choice task. If the range of observed probabilities for the expensive alternative covers the [0,1] range, this gives an indication that the estimated VTT distribution is supported by the data. The results of this process are reported in Table 1. A word of caution is required here. The models were estimated jointly across all three games, but this simple test is only possible for SP1 (given the simple two-attribute trade-off and log-WTP space setup). There is thus a possibility that values higher than those supported in SP1 are supported in SP2 and SP3, as mentioned in the second point above. Nevertheless, the ranges in probabilities revealed by this test support our argument that the trade-offs presented were wide enough, with the important upper tail of the distribution being observed beyond the 98.5% fractile in all six samples, and beyond the 99.5% fractile in four of the six samples. For further interpretation of this test, see Fosgerau et al. (2007a).

Table 1: Range of estimated probabilities for expensive alternative in SP1

	Range of probabilities for expensive option in SP1 using final model	
	Car	**Rail**
Commute	[0.0167,0.9874]	[0.0007,0.9984]
Employees’ business	[0.0047,0.9964]	[0.0008,0.9979]
Other non-work	[0.0118,0.9846]	[0.0237,0.9989]

For bus and ‘other PT’, we were additionally able to estimate random scale heterogeneity across respondents. A possible reason for our inability to estimate such scale heterogeneity also for car and rail is that, with the longer trips for these modes, the main source of scale heterogeneity would be distance based, and this is already captured by the multiplicative models.

In particular, to introduce random scale heterogeneity into the bus and ‘other PT’ games, we used:

$$\mu_{SPj} = e^{\mu_{log\left(\mu_{SPj}\right)}+\sigma_{log(\mu)}\xi_N} \tag{33}$$

where ξ_N is a standard normal random variate, and where the resulting scale parameters now follow a lognormal distribution.

3.6. Overview and additional points

The discussions in the preceding subsections have looked at the individual model components in turn. While a joint estimation was used across games, differences arise in the specification of the individual components for each game. An overview of the model specification is given in Table 2, linking back to the individual sub-sections and equations therein.

Table 2: Overview of model specification

	SP1	SP2	SP3
Underlying error structure	log-wtp space eqns (1) & (2)	multiplicative WTP space eqn (6) for SP1/SP3 and car SP2, eqn (10) for other SP2	
Implementation of dBF (size and sign effects)	eqns (18) & (19)	eqn (21), using specification (22) for car SP3 and (23) for PT SP3	
Specification of deterministic heterogeneity	e.g. eqns (25) & (26)		
Choice probabilities before mixing	eqn (3)	eqn (7)	
Joint choice probability before mixing	eqn (24)		
Specification of random heterogeneity	$\theta_0 = e^{r_u}, where\ r_u \sim U[a, a+b]$		
Choice probability with mixing	eqn (30)		
Value of Travel Time (for game *x*)	$\theta_0^{\kappa_{SPx,VTT}} \zeta_{SPx,VTT} \prod_m z_m^{\lambda_m} \prod_n \zeta_n^{z_n} \cdot \lvert\Delta t\rvert^{\kappa_{SPx,VTT}-1}$		

Our joint estimation allows for differences in valuations across the individual games but assumes that the underlying heterogeneity and socio-demographic effects are proportional to these differences. This is of course a strong assumption, especially in relation to SP2, as the correlation between the value of reliability and the value of time may not be as strong as say the correlation between the values of time in different travel conditions. To test the impact of our joint specification, we performed a simple test where we ran the specification of the final car commute model on the SP1 data only, and this led to a change in the VTT by only 3.4%, suggesting that the use of game specific multipliers was sufficient to avoid any potential biasing impact of individual games, e.g. SP2, on the overall values. This justifies the use of a joint modelling approach, given the benefits in terms of covariates.

4. Estimation results for joint models

We now proceed with the presentation of the estimation results, where, for space reasons, the presentation in this paper focusses on the results for car and rail travellers. In general, parameters with a low level of statistical significance were removed from the model, with the exception of a number of key multipliers where we did not wish to impose equality with base categories. Parameters that did not have a significant estimate across any of the three purposes in estimation are not shown in the tables which follow; this applies for example to a large number of the dBF parameters, in addition of course to many of the covariates tested, whether traveller, trip or design related.

4.1. Results for car

The results for the car models for the three travel purposes are presented in Table 3, and can be summarised as follows:

parameters of base θ_0 distribution

- The models retrieve significant random heterogeneity for all three purpose segments as can be seen by the statistically significant and large estimates for $b_{\log(\theta_0)}$, which gives the range of the underlying uniform distributions, while $a_{\log(\theta_0)}$ gives the minimum value of the underlying distribution, so that its exponential gives the minimum value of θ_0.

game-specific θ_0 multipliers

- As the game-specific multipliers cannot directly be interpreted as the differences in the valuations across games (cf. earlier discussion) analysis of the differences across games needs to happen at the implementation stage, on the basis of specific assumptions about Δt (see Section 5).

key elasticities

- Significant positive income elasticities on the various VTT measures are obtained across all three purposes; these are highest for other non-work, and lowest for employees' business.
- There are significant positive elasticities on the valuations in relation to reference cost and negative elasticities in relation to reference time, with the cost elasticities generally larger in absolute value than the time elasticities, implying that the VTT is higher for longer distance trips (which tend to be more expensive and take longer). The elasticities are strongest for other non-work, and weakest for business. These time and cost elasticities can be related

to the damping effects on longer trips, where heterogeneity in valuations implies that sensitivity to both time and cost diminishes on longer trips, so cost increases would *increase* VTT, while time increases would *reduce* VTT.

- A significant distance elasticity is observed only for employees' business, which is positive, leading to higher valuations on longer trips even with time and cost being held constant. The 2003 study specified the cost elasticity as a proxy for distance, as no distance information had been collected in the survey, and hence could not attempt to distinguish between the two effects, but it has also been argued before that impacts of cost and time make more behavioural sense than a pure distance impact (Daly, 2010). These effects can be plausibly related to self-selection by travellers for the journeys concerned.

traveller covariates

- There are differences in VTT for travellers who do not report income, with separate effects for different groups.
- Female commuters show higher valuations, while younger travellers, all else being equal, also have higher valuations for commute and for other non-work, where, for the latter, this captures both of the lowest two age categories.
- There are lower valuations for households with two or more adults in the other non-work segment.
- Households owning at least one car have higher valuations for other non-work trips, while households with two or more motorcycles have much lower valuations in the same segment.
- Self-employed commuters have higher valuations, as do commuters where travel costs are paid by the company.
- For employees' business trips, the valuations are higher if the company buys savings come what may, and lower if it does not buy time savings.
- For employees' business trips, valuations are lower if self-employed, and then lower still if self-employed costs are not covered.

trip covariates

- Valuations are higher on other non-work trips with at least one night away from home.
- Commuters travelling with others have lower valuations.
- Commuters have lower valuations when driving on rural roads.
- The impact of congestion during the reference trip is only evident for commuters and other non-work, and the statistical significance of the effects is low, albeit that we see higher valuations for respondents with more congestion on their reference trips.
- For employees' business travel with a London origin and destination, we observe higher valuations, although the statistical significance of the effect is low.

design covariates

- We observe lower valuations when the cheap option is presented on the left for both SP1 and SP3 for commute and other non-work, where the effects are less strong and not as significant for SP3 as for SP1, or for other non-work as opposed to commute.
- Valuations are lower in SP1 for employees' business if time is shown above cost.
- The scale (i.e. the inverse of the variance of the random error term) for SP2 is lower for commuters if SP2 is shown before SP3, and higher for employees' business.

scale parameters

- The values for the scale parameters should not be compared between SP1 and SP2/SP3 given the different modelling approach that was used for SP1.
- No specific insights could be obtained from comparing the relative values of μ_{SP2} and μ_{SP3} across purposes.

dBF parameters

- There are sign effects (gain-loss asymmetry) for time in SP1 and SP2 for commuters and SP1 for other non-work, while we note sign effects for cost for commuters in SP1 and SP3, and in SP2 for employees' business and other non-work. This means that, in the contexts noted, valuations vary depending on whether there is a time/money gain or loss.
- We observe size effects for time in SP1 and SP2 across all three purposes, and for cost in SP1 for employees' business and other non-work. This means that, in the contexts noted, valuations vary by the size of the time/cost change away from the reference values.
- There is asymmetric damping for time in SP1 across all three purposes, and in SP2 for employees' business and other non-work, while, for cost, asymmetric damping is observed only for employees' business and then only in SP3. Together with the findings for the sign effects, this could suggest a more binding constraint on time than on money, at least in the short term.

Table 3: estimation results for joint car models

	Commute		Employees' business		Other non-work	
Respondents	922		917		977	
Observations	13,830		13,755		14,655	
Final log-likelihood	-7,332.67		-6,933.43		-7,585.74	
adjusted ρ^2	0.23		0.27		0.25	
parameters of base θ_0 distribution	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**
$a_{\log(\theta_0)}$	-0.3559	-1.74	0.5150	3.61	-0.8840	-2.65
$b_{\log(\theta_0)}$	3.7060	15.62	3.3727	18.31	3.7141	19.16
game specific θ_0 multipliers	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**
SP1 travel time	1	-	1	-	1	-
SP2 travel time	1.5988	4.05	1.1396	0.82	2.1875	5.52
SP2 std dev of travel time	0.5803	-4.75	0.8765	-1.04	0.8118	-1.48
SP3 free flow	0.6968	-2.26	0.5718	-4.54	0.5008	-4.43
SP3 light congestion	0.9770	-0.14	0.9206	-0.74	0.8801	-0.90
SP3 heavy congestion	1.8557	2.98	1.7076	4.23	1.9955	4.05
key elasticities	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**
income elasticity (λ_{income})	0.5797	6.10	0.3003	3.64	0.6819	7.76
distance elasticity ($\lambda_{\mathrm{distance}}$)	0	-	0.2390	3.41	0	-
cost elasticity (λ_{cost})	0.6790	3.70	0.4511	2.63	1.0492	6.56
time elasticity (λ_{time})	-0.6241	-2.62	-0.4538	-2.29	-0.9273	-4.72
traveller covariates (multipliers on θ unless stated)	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**
unstated income ($\zeta_{income\ not\ stated}$)	2.4775	0.65	0.5034	-2.41	1.0117	0.03
unknown income ($\zeta_{income\ unknown}$)	1.4264	1.16	9.3098	2.90	0.2998	-5.89
refused income ($\zeta_{income\ refused}$)	0.7697	-1.30	0.5812	-1.43	0.8644	-0.77
female (base=male)	1.3674	2.26				

aged 17-29 (base=30+)	1.3645	1.76				
aged 17-39 (base=40+)					1.4530	2.52
household with 2+ adults (base=1 or no adults)					0.6980	-3.47
1+ car owned (base=no cars)					2.6826	1.91
2+ motorcycles owned (base=1 or 0 motorcycles)					0.4668	-1.65
Self-employed (base=any other)	1.6669	1.97				
Travel costs paid by company (base=respondent or other paid)	2.2194	3.09				
Company would buy savings come what may (base=buys if benefits>costs, or unknown)			1.3044	1.34		
Company would not buy time savings (base=buys if benefits>costs, or unknown)			0.4435	-9.51		
Self-employed costs not covered (base=costs covered)			0.5629	-3.04		
Self-employed (base=paid employment)			0.6767	-2.56		
trip covariates (multipliers on θ unless stated)	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**
1+ nights away (base=day return)					1.5522	2.14
travelling with others (base=travelling alone)	0.6690	-3.37				
driving on rural roads (base=urban or motorway)	0.8119	-1.31				
light congestion (base=free flow)	1.4025	1.57			1.3554	1.51
heavy congestion (base=free flow)	1.5604	1.78			1.4621	1.57
trip with London base origin & destination (base=any other)			1.7530	1.42		
design covariates (multipliers on θ unless stated)	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**
SP1 cheap option on left (multiplicative effects coding)	0.8842	-3.26			0.9259	-2.00
SP3 cheap option on left (multiplicative effects coding)	0.9282	-1.37			0.9537	-0.85
SP1 time shown above cost (multiplicative effects coding)			0.8878	-2.52		
SP2 scale (μ_{SP2}) multiplier if SP2 before SP3 (multipl. effects coding)	0.8938	-2.63	1.1531	2.52		
scale parameters	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**
μ_{SP1}	1.1975	14.71	1.7354	16.90	1.3014	16.53
μ_{SP2}	7.7383	18.05	6.3695	10.95	7.5389	16.92
μ_{SP3}	5.6636	14.65	7.2603	16.16	5.9410	17.07
dBF parameters	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**

$\beta_{t,SP1}$	-0.4000	-3.64	-0.1141	-1.61	-0.1366	-1.91
$\beta_{t,SP2}$	-0.1564	-2.84	-0.4487	-5.10	-0.2435	-4.96
$\beta_{c,SP1}$			0.1013	1.83	0.1032	1.78
$\gamma_{t,SP1}$	-0.2127	-3.52	-0.1293	-3.58	-0.1075	-2.75
$\gamma_{t,SP2}$			-0.0627	-1.94	-0.0606	-1.79
$\gamma_{c,SP3}$			-0.1581	-2.11		
$\eta_{t,SP1}$	0.2573	4.34			0.2237	4.20
$\eta_{t,SP2}$	0.0874	1.43				
$\eta_{c,SP1}$	0.1267	2.18				
$\eta_{c,SP2}$			0.1959	2.15	0.2244	2.88
$\eta_{c,SP3}$	0.2771	1.51				

4.2. Results for rail

The results for the rail models are presented in Table 4 and can be summarised as follows:

parameters of base θ_0 distribution

- The models retrieve significant random heterogeneity for all three purposes.

game specific θ_0 multipliers

- The multipliers (relative to the time measured in SP1) are all of the correct sign, with the negative multipliers for early arrival relating to a reduction in trip time.
- We can draw some conclusions in relation to the crowding multipliers in SP3 as the same dBF parameters apply to all these multipliers, where we note a monotonic increase in sensitivity to different levels of crowding for both seated and standing passengers. Across purposes however, the sensitivity to highest crowding level for seated passengers is higher than the sensitivity to the lowest crowding level for standing passengers, and in fact for the two lowest levels of crowding for standing passengers for commute and other non-work. For other non-work, the sensitivity to the two lowest levels of crowding for seated passengers is constant.

key elasticities

- Significant positive income elasticities on the various VTT measures are obtained across all three purposes.
- A distance elasticity is observed only for employees' business, which is positive but not highly significant.
- There are significant positive elasticities on the valuations in relation to cost, and negative elasticities in relation to time, implying that the VTT is higher for longer distance trips.

traveller covariates

- A diverse picture emerges for the various multipliers for respondents without income information, where the only highly significant effect is a much lower set of valuations for commuters with unstated income.
- Female respondents on other non-work trips have lower valuations.
- There are higher valuations for households with three or more children in the other non-work segment, and lower valuations for households with three or more adults in the commute segment.
- For commute trips, valuations are higher if costs are paid by the company or any other party, while, for other non-work trips, they are higher if costs are paid by the company.
- For employees' business trips, the valuations are higher if the company buys savings come what may, and lower if it does not buy time savings or if the policy is unknown to the respondent.
- For employees' business trips, valuations are lower if self-employed, especially for blue collar.

trip covariates

- Valuations are lower for commuters on trips with overnight stays, higher for other non-work for one-night return trips, and lower for employees' business on trips with multiple nights away from home.

- Lower frequency leads to lower valuations for other non-work (if less than daily) and commute (if less than monthly).
- Valuations are lower for commute and employees' business for one-way trips.
- Valuations are lower for weekend travel for other non-work.
- Valuations are lower for travellers without a reserved seat for other non-work.
- For employees' business travel with a London origin and destination, we observe higher valuations.

design covariates

- We observe lower valuations when the cheap option is presented on the left for SP1 for other non-work.

SP2 specific effects

- The negative values for α show risk seeking behaviour across all purposes.
- There is an overall preference for alternatives with constant travel times, i.e. no variability.

scale parameters

- The values for the scale parameters should not be compared between SP1 and SP2/SP3 given the different modelling approach that was used for SP1.

dBF parameters

- There are sign effects (gain-loss asymmetry) for time in SP1 for commuters and employees' business, for time in SP2 for commute and other non-work, for cost in SP2 for commute and other non-work, and for cost for commute in SP3.
- We observe size effects for time in SP1 and SP3 across all three purposes, and for cost in SP1 for employees' business and in SP2 for all purposes.
- There is asymmetric damping for time in SP1 for commute and other non-work, and for time in SP3 and cost in SP2 for commute and employees' business. Together with the findings for the sign effects, this could suggest a more binding constraint on time than on money, at least in the short term.

Table 4: estimation results for joint rail models

	Commute		Employees' business		Other non-work	
Respondents	847		945		996	
Observations	12,340		13,390		14.275	
Final log-likelihood	-6,016.62		-6,903.61		-7,371.27	
adjusted ρ^2	0.29		0.25		0.25	
parameters of base θ_0 distribution	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**
$a_{\log(\theta_0)}$	0.4305	4.15	0.6025	5.21	0.8655	4.55
$b_{\log(\theta_0)}$	2.7356	18.92	2.6219	20.66	2.8442	22.52
game specific θ_0 multipliers	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**
SP1 travel time	1	-	1	-	1	-
SP2 travel time	1.2356	2.58	1.8071	4.45	1.0877	0.70
SP2 early delay	-2.1925	-2.83 (vs -1)	-2.7967	-3.28 (vs -1)	-2.5482	-3.49 (vs -1)
SP2 late delay	3.5360	5.07	4.9920	5.36	3.4967	6.10
SP3 seated with 50% Load Factor	0.7033	-3.07	0.8509	-1.20	0.7336	-2.42
SP3 seated with 75% Load Factor	0.7621	-2.45	0.8618	-1.08		
SP3 seated with 100% Load Factor	0.9695	-0.29	1.1280	0.89	1.0242	0.19
SP3 seated with 1 pass standing per m^2	1.0543	0.50	1.2790	1.83	1.1642	1.17
SP3 seated with 3 pass standing per m^2	1.2704	2.24	1.5289	3.13	1.4280	2.69
SP3 standing with 0.5 pass per m^2	1.1216	0.97	1.4506	2.40	1.2417	1.51
SP3 standing with 1 pass per m^2	1.1574	1.19	1.5612	2.79	1.2962	1.69
SP3 standing with 2 pass per m^2	1.2750	1.88	1.7642	3.20	1.6055	3.04
SP3 standing with 3 pass per m^2	1.5246	3.25	1.8148	3.00	1.8298	3.66
SP3 standing with 4 pass per m^2	1.8026	4.49	2.2878	4.05	2.2197	4.74
key elasticities	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**
income elasticity (λ_{income})	0.2979	4.44	0.3566	5.69	0.2936	6.10
distance elasticity ($\lambda_{\mathrm{distance}}$)	0	-	0.0585	1.15	0	-
cost elasticity (λ_{cost})	0.6640	7.82	0.7428	12.67	0.5983	10.35
time elasticity (λ_{time})	-0.2753	-2.49	-0.3479	-3.87	-0.5406	-7.20

traveller covariates (multipliers on θ unless stated)	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**
unstated income ($\zeta_{income\ not\ stated}$)	0.3871	-5.58	1.0728	0.24	0.6251	-1.50
unknown income ($\zeta_{income\ unknown}$)	1.5947	0.82	1	-	1.1516	0.52
refused income ($\zeta_{income\ refused}$)	0.6543	-1.62	2.4822	0.93	1.1719	0.59
female (base=male)					0.8429	-2.37
household with 3+ children (base=2 or fewer children)					1.6202	2.10
household with 3+ adults (base=2 or fewer adults)	0.8539	-1.64				
Travel costs paid by company or other (base=respondent paid)	2.0689	4.16				
Travel costs paid by company (base=respondent or other paid)					1.6473	2.39
Company would buy savings come what may (base=buys if benefits>costs)			1.4692	2.09		
Company policy on savings unknown (base=buys if benefits>costs)			0.5616	-2.62		
Company would not buy time savings (base=buys if benefits>costs)			0.3504	-19.00		
Self-employed briefcase (base=paid employment)			0.5489	-5.82		
Self-employed blue collar (base=paid employment)			0.3612	-6.64		
trip covariates (multipliers on θ unless stated)	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**
1+ nights away (base=day return)	0.5676	-5.22				
1 night away (base=day return or 2+ nights away)					1.4008	2.40
2+ nights away (base=day return or 1 night away)			0.7714	-2.29		
frequency less than once per day (base=daily)					0.6477	-2.75
frequency less than once per month (base=1 or more times per month)	0.7227	-2.60				
one-way trip (base=return trip)	0.6023	-2.80	0.6996	-2.04		
weekend travel (base=weekday travel)					0.6439	-3.30
no reserved seat (base=reserved seat)					0.8514	-1.90
trip with London base origin & destination (base=any other)			1.9416	2.86		
design covariates (multipliers on θ unless stated)	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**
SP1 cheap option on left (multiplicative effects coding)					0.9490	-2.13
SP2 specific effects	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**
risk averseness parameter (α)	-0.0395	-3.11	-0.0070	-2.48	-0.0279	-3.81

	est.	rob t-rat (0)	est.	rob t-rat (0)	est.	rob t-rat (0)
constant for alternatives with zero variability (expressed in £)	0.4465	4.60	0.3814	2.12	0.5682	5.44
scale parameters	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**
μ_{SP1}	1.8210	18.90	2.1018	20.06	1.8857	21.80
μ_{SP2}	7.2521	9.91	10.8840	10.40	6.9026	12.09
μ_{SP3}	6.5578	13.48	7.1558	11.31	6.6187	12.93
dBF parameters	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**	**est.**	**rob t-rat (0)**
$\beta_{t,SP1}$	-0.2137	-3.31	-0.1462	-3.62	-0.1327	-2.79
$\beta_{t,SP3}$	-0.2418	-2.51	-0.1421	-3.14	-0.1861	-3.72
$\beta_{c,SP1}$			0.0683	1.75		
$\beta_{c,SP2}$	0.2478	3.78	0.1351	1.95	0.1587	2.84
$\gamma_{t,SP1}$	-0.1093	-2.39			-0.1364	-5.19
$\gamma_{t,SP3}$	-0.1216	-2.00	-0.1016	-3.48		
$\gamma_{c,SP2}$	0.3456	3.40	-0.2641	-3.61		
$\eta_{t,SP1}$	0.0951	1.63	0.1109	3.87		
$\eta_{t,SP2}$	0.1408	2.18			0.2858	4.40
$\eta_{c,SP2}$	0.2058	2.01			0.2319	1.54
$\eta_{c,SP3}$	0.1379	1.90				

5. Translation into values for appraisal

5.1. Sample enumeration work and error calculations

The eleven estimated behavioural models (i.e. one for each available mode-purpose combination, noting that business travel for bus was excluded from our scope) relate to the estimation sample alone. Despite its broad representativeness, reweighting is required to produce fully representative values. To this end, an R-based Implementation Tool was programmed matching each observed trip in the English National Travel Survey (NTS) to its corresponding behavioural model. The NTS is an established series of household surveys of personal travel in Great Britain, designed to track the long-term development of trends in travel. NTS data is collected via interviews with people in their homes and a diary that they keep for a week to record their travel. The NTS covers travel by all age groups, including children. In each year, diary data was collected from over 7,700 households, covering over 18,000 individuals. We used NTS data for the years 2010-2012, where fares and incomes have been adjusted to 2014 prices and values using CPI for fares and CPI and real income growth for incomes. The response rates from the NTS survey were 60% in 2010 and 61% in 2011 and 2012.

For each trip in the NTS, the travel cost and travel time were identified in combination with the corresponding socio-economic characteristics, and appropriate values were then calculated for each of the different valuations from our models, along with accompanying error measures.

Trip specific mean and variance in VTT

The implemented models, using a mixed logit formulation, do not produce a single estimate of VTT for a particular trip but rather a distribution. For the Implementation Tool, the mean and variance of the distribution are used to represent the distribution. Based on (28) and the estimated parameters a and b of the log-uniform distribution, the mean and variance of the value of time for trip i (using SP_x) are:

$$E(VTT_i) = \exp(\kappa_{SPx}a)\,\frac{\exp(\kappa_{SPx}b)-1}{\kappa_{SPx}b}\,\zeta_{SPx,VTT}\prod_m z_{mi}^{\lambda_m}\prod_n \zeta_n^{z_{ni}}\,|\,\Delta\,t|^{\kappa_{SPx}-1} \quad (34)$$

$$Var(VTT_i) = \exp(2\kappa_{SPx}a)\left[\frac{\exp(2\kappa_{SPx}b)-1}{2\kappa_{SPx}b} - \frac{(\exp(\kappa_{SPx}b)-1)^2}{\kappa_{SPx}{}^2b^2}\right]\left(\zeta_{SPx,VTT}\prod_m z_{mi}^{\lambda_m}\prod_n \zeta_n^{z_{ni}}\,|\,\Delta\,t|^{\kappa_{SPx}-1}\right)^2 \quad (35)$$

Note that the above measure of variance does not relate to estimation uncertainty and sampling error, but only to unobserved heterogeneity in preferences across the population[9].

Sample enumeration

VTT estimates were calculated by sample enumeration using a sample of trips drawn from the NTS. The NTS trips are weighted by expansion factors provided

[9] To explain the rationale for the within-record variance, we note that for a given person we do not know where that person is positioned on the estimated distribution of unobserved preference heterogeneity. Hence the corresponding variance in (35) denotes this uncertainty related to a specific record.

with the NTS survey[10] and the trips can additionally be distance weighted, so that the VTT from sample enumeration represents the VTT for an average kilometre.

This approach implies calculating the distance-weighted VTT for a given population segment given values for *E(VTT)* and Δt:

$$\overline{VTT_S}(\Delta t) = \frac{\sum_{i \in S} w_i l_i E\left(VTT_i(\Delta t)\right)}{\sum_{i \in S} w_i l_i} \tag{36}$$

where we take the sum over every NTS trip i in the segment S, with w_i being the NTS expansion weight for the trip and l_i being the relevant trip length. The expected VTT depends on the covariates of trip i ensuring variation across records.

In discussing the variation of VTT, we need to distinguish carefully between the variation of VTT in the population, which we describe in our model, and the error arising in the model because the parameters are estimated with error. First, we focus on the population variance T of VTT, which comprises the within-record variance T_1 and the between-record variance T_2, which are independent, so that $T=T_1+T_2$. The within-record variance is generated by the mixed logit model and was already defined above (35). The between-record variance can be calculated as:

$$T_2 = \frac{\sum_{i \in S} w_i l_i \left(VTT_i(\Delta t) - \overline{VTT}_S(\Delta t)\right)^2}{\sum_{i \in S} w_i l_i} = \frac{\sum_{i \in S} w_i l_i \left(VTT_i(\Delta t)\right)^2}{\sum_{i \in S} w_i l_i} - \left(\overline{VTT_S}(\Delta t)\right)^2 \tag{37}$$

This is a relatively straightforward calculation when the mean is also being accumulated.

Error in the mean

To calculate error in the mean VTT estimate for a segment, we apply the 'delta method' for the variance of a function of random variables, which can be shown to be in some respects optimal when applied to maximum likelihood estimates (Daly et al. 2012). The error in $\overline{VTT}_S$ is calculated by:

$$var(\overline{VTT}_S) = \Phi' \Psi \Phi \tag{38}$$

where Φ is the vector of first derivatives of $\overline{VTT}_S$ with respect to the estimated parameters and Ψ is the covariance matrix of those parameters. Given the formulation of $\overline{VTT}_S$ it is clear that:

$$\Phi = \frac{\sum_{i \in S} w_i l_i \phi_i}{\sum_{i \in S} w_i l_i} \tag{39}$$

where ϕ_i is the vector of first derivatives of VTT for the specific record i. For calculation, therefore, we need to accumulate the components of Φ at the same time, and with the same weights, as we accumulate the VTT itself.

[10] The NTS expansion factors weight a specific trip based on the frequency of reporting long and short distance trips in the NTS travel diaries relative to a nationally representative sample whilst accounting for the drop-off recording during the week in which the travel diary is registered. Additionally, it accounts for non-response (incomplete surveys) at the household level, based on analyses that are not available to us. The applied weights are those as recommended by the official NTS documentation.

Error in the NTS sample: bootstrapping

The NTS sample used in this study is large so that it should give a reliable picture of the total population. Further, weights are provided with the data that should correct for several biases, such as differential response by specific population groups. However, it remains a sample survey and is thus subject to error. Moreover, some of the segments of interest are small fractions of the total population, so that for these segments the sample error may be larger. It is, therefore, useful to develop methods for estimating the error arising because of the sample nature of the NTS data. The method used for this calculation was the 'bootstrap'.

The bootstrap method works by constructing different samples from the original sample. This is done by drawing samples of the original size from the original sample, with replacement, so that some records may be sampled several times and others not at all. The well-researched literature on the bootstrap method assures us that the variation across these samples gives a good and unbiased representation of the true variation due to sampling.

5.2. Assumptions made for Δt

As already discussed in Section 3, the calculation of the VTT measures from our models is dependent on an assumption relating to Δt (see e.g. (15)). Different VTT measures will be obtained with different sizes of time changes. This is reflected in the formulae in the present section once again also being a function of Δt. In appraisal, an overriding consideration is that we should obey basic laws of 'adding up': the incremental value of (A minus B) plus the incremental value of (B minus C) should equal the incremental value of (A minus C). Thus, there is a strong appraisal argument for the use of a constant unit value, and we would need equally strong counter arguments to move away from this position, which appears to be taken by most governments (Daly *et al.*, 2014). Nevertheless, given the dependence of VTT on Δt, there is still the need to take a view on what value to use.

Based on the variation of VTT with Δt, the previous British study (in effect[11]) made the calculation of VTT for appraisal based on a Δt of 11 minutes (Mackie et al., 2003). The Danish VTT study (Fosgerau et al. 2007a) suggested different values of Δt for different modes, though for most modes the calculated VTT was stable for Δt between 10-20 min. Based on this, they decided that 10 minutes was 'reasonable' for all modes. The Swedish VTT study (Börjesson and Eliasson 2014) also chose varying Δt values, especially between long and short distances: in the event, they decided to use 15 minutes for regional trips and 20 minutes for long distance, though these were essentially arbitrary decisions based on what was felt to be 'reasonable'. In Norway the VTT for short distance modes turned out to be rather stable at Δt =10 min, but not for long distance modes. They used a Δt of 10 minutes for short distance travel and 15 minutes for all long-distance modes, following similar reasoning to the British, Danish and Swedish studies.

[11] Although the so-called 'perception' function used in that study differed from the dBF approach used here, the same issues apply.

Good practice would suggest that the estimated behavioural models should only be applied within the range of covariate values over which they have been estimated. The range of Δt presented in our own SP data tended to be on the low side (particularly for bus and 'other PT', where more than 80% of the values were less than 10 minutes), but for car and rail a reasonable proportion (more than 15%) were greater than 20 minutes. Based on these distributions, it was decided with the Department that a value of Δt of 10 minutes is defensible as a basis for appraisal values. The values presented in the official British VTT report are therefore all based on the assumption that Δt = 10 minutes.

Table 5 presents the outputs of the Implementation Tool at Δt = 10 minutes for each mode-purpose combination, focussing for now on SP1. The results confirm general patterns that VTT is highest for the car and rail alternatives, but lower for bus and other forms of public transport. This distinction is most apparent for the business trips (EB) and less prevalent for other non-work related trips. Besides there being significant variation in the VTT across mode-purpose combinations, there is also significant variation in the VTT within the populations of interest. With the application of log-uniform (and more so with lognormal) densities in the mixed logit model, we see that the population standard deviation exceeds the mean VTT. This is a direct consequence of the right-skewness of the log-uniform distribution, which also drives the discrepancy between the mean and median in the VTT distribution.

Uncertainty surrounding the weighted population mean VTT estimates is primarily driven by estimation uncertainty. Table 5 shows that controlling for potential sampling error in the NTS data increases the error only slightly. However, the standard errors are still relatively small and in most cases between 10-15% of the mean value.

Table 5: Output of the Implementation Tool for SP1 based on Δt = 10 minutes. (VTT estimates in £/hour, 2014 perceived prices)

Commute

	Mean	Pop. Stdev	St. err. mean	St. err. mean + bootstrap
Car	11.695	18.654	1.969	1.970
Bus	3.148	3.260	0.458	0.458
OtherPT	6.348	6.472	0.874	0.876
Rail	12.423	19.284	0.866	0.898

EB

	Mean	Pop. Stdev	St. err. mean	St. err. mean + bootstrap
Car	16.738	22.123	1.953	1.959
Bus	NA	NA	NA	NA
OtherPT	8.328	10.635	1.305	1.330
Rail	27.609	38.209	2.388	2.599

OtherNW

	Mean	Pop. Stdev	St. err. mean	St. err. mean + bootstrap
Car	4.910	9.196	1.756	1.756
Bus	3.257	4.276	0.410	0.411

OtherPT	5.232	5.576	0.550	0.553
Rail	8.675	10.472	0.605	0.615

5.3. Differences across valuations

The behavioural modelling work conducted in this study provides a rich set of results, with a number of different valuations for travel time obtained across (in general) three different SP games. This represents an extension beyond traditional European VTT work, which has relied solely on SP1 results when producing values for appraisal, but introduces a need to check whether there is substantial evidence of differences across the individual games.

We now consider the behaviour of the VTT outputs for across the different SP games and the corresponding sensitivity to the selection of Δt. Figure 1 presents the different valuations obtained for car travellers across the different games and across all three purposes. We show these calculations across a range of Δt values.

SP1 and SP2 were the only situations in which we could detect behavioural effects associated with Δt. Accordingly, we see constant values for the valuation of standard deviation of time (SP2sd) and the three valuations from SP3. The expected ordering for SP3, with the VTT in heavy congestion (VTT_{hc}) being highest, followed by the VTT for travel in light congestion (VTT_{lc}) and in free flow conditions (VTT_{ff}). The graphs clearly show the important role of Δt and the corresponding interpretation of the SP1 (and SP2) estimates. At $\Delta t=1$ the SP1 value is very close (or even below) the VTT_{ff} estimate. The VTT in SP1 and SP2 then increases rapidly, but less steeply in the range between $\Delta t=10$ and $\Delta t=20$.

Whilst the SP1 valuation is typically positioned between the VTT_{hc} and VTT_{lc} values, this changes depending on the assumption made for Δt, with the SP1 value eventually exceeding the high congestion value from SP3. As such, it is extremely hard to justify interpreting the corresponding SP1 value as one relating to trips with ‘medium’ congestion.

Similarly, the fact that the SP2 valuation differs from the SP1 valuation casts further doubt as to what type of time the valuations in SP1 relate to. For this reason, many studies around the world have moved away from approaches not describing travel time as relating to specific conditions and instead rely on values from games of the SP3 type. One can then for example also produce also a weighted VTT accounting for the real-world mix of traffic conditions (which may differ from that in the sample) or look at how that weighted VTT might evolve if traffic conditions change. Of course, one could find a value for Δt that would mean that the VTT from SP1 corresponds to a weighted value from SP3, but it would then not be possible to adjust this to different traffic conditions but would remain a function of the (latent) traffic conditions that respondents consider when answering SP1 type questions. The fact that SP3 seems not to be affected by reference dependence effects to the same extent, makes its values easier to use (not requiring an assumption for Δt) and also poses the question of whether reference dependence effects have been amplified by the survey setting. Additionally, the use of SP3 results in appraisal gives greater flexibility in updating values over time as a function of changes in congestion. Similar patterns to these car travel results were also observed for the other modes.

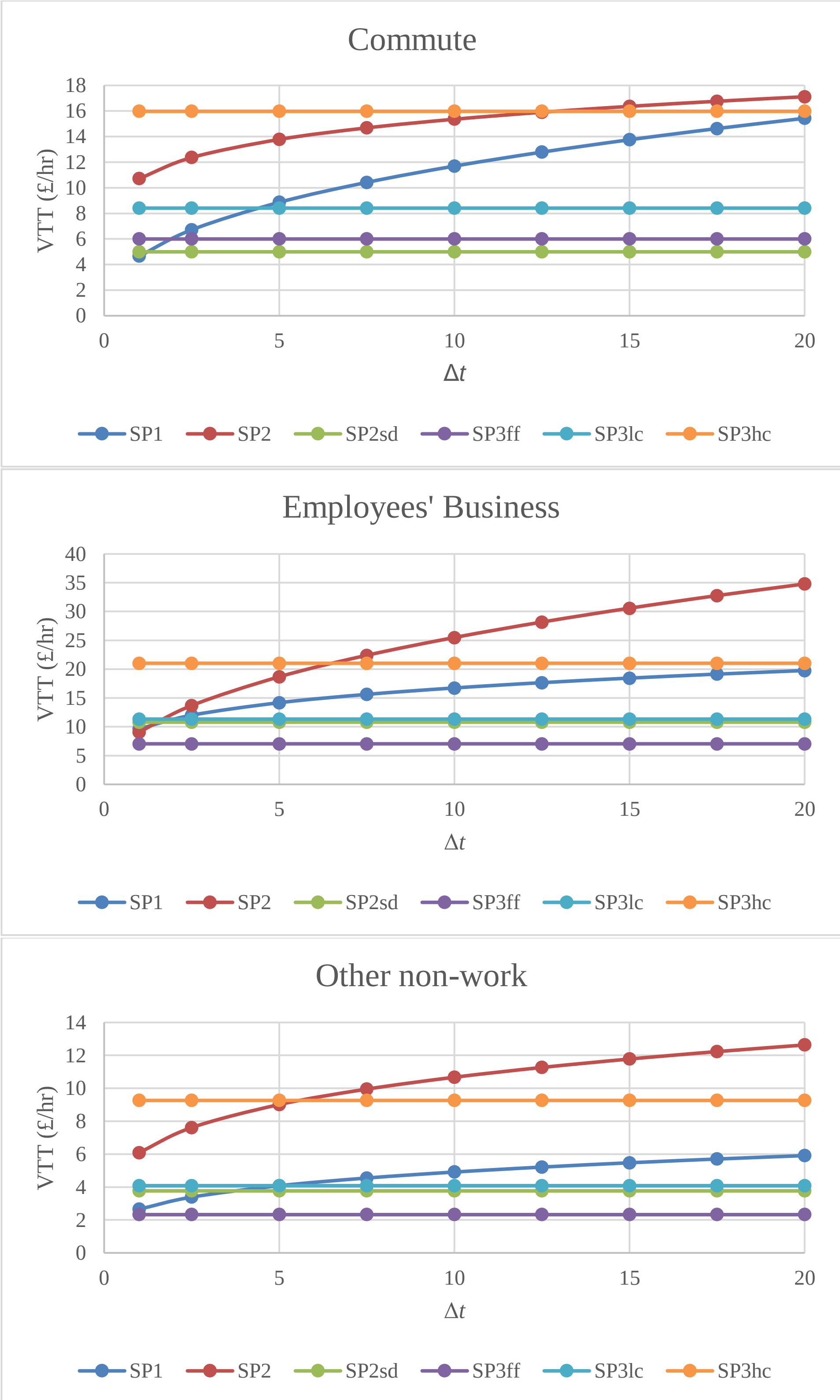

Figure 1: VTT measures across different games and Δt assumptions for car travel

6. Conclusions

This paper has given an overview of the data collection and estimation work carried out for the 2014/2015 national value of time study in Great Britain. The study used data from Stated Preference (SP) surveys on which we estimated advanced discrete choice models to produce valuations for different components of travel time and journey conditions. Our study used three games, which considered different trade-offs, namely: SP1 (time vs. money), SP2 (time vs. money vs. reliability), and SP3 (time vs. money vs. crowding/congestion).

The modelling of this data was developed in a systematic fashion, whereby alternative model specifications were tested in relation to:

- time and cost gains, losses and size effects;
- person characteristics such as age, gender, employment status, household composition and income; and
- trip characteristics such as mode, purpose, distance and geography.

Our modelling work made use of state-of-the-art approaches and made a number of departures from current practice in the UK. We additionally made some improvements over the leading work from the recent Scandinavian studies, for example in terms of reference dependence in games other than SP1. Our contributions to the methodological framework are three-fold. Firstly, we extend the De Borger and Fosgerau (2008) framework for reference dependence and asymmetry to the case of games with more than two attributes using a multiplicative error specification. Secondly, we allow for joint estimation across different SP games in the presence of differences in these size and sign effects, which requires an alternative specification of deterministic heterogeneity. Thirdly, we seem to offer the first large scale application of the log-uniform distribution in choice modelling, thereby minimising complications associated with the right tail of the VTT distribution across the population.

The use of a joint modelling framework across all the three SP games allowed us to examine the differences between individual valuations across games. This joint estimation yielded key insights, for example showing the extent of VTT increases with road congestion (for car and bus), as well as with the level of crowding (for all PT modes). It also showed differences across modes and across purposes as to what type of trip conditions in SP3 the values from SP1 relate to. This raises questions as to the interpretation of SP1, in terms of the traffic conditions perceived by respondents when making their choices. Overall, SP1 also shows more prevalent size and sign effects than SP2 and SP3, provoking the question of whether simpler settings increase the scope for reference dependence. Finally, the values coming out of SP2 diverge from those of SP1 and SP3, and this is likely to be a behavioural impact in terms of how respondents react to variability in trip times. In general, our findings arguably call for the use of more detailed scenarios (in terms of moving away from simple time-money trade-offs), as is done in many other countries, while the differences between SP2 and SP3 results potentially call for a joint treatment of reliability and quality.

More generally, the methodological developments reported here further extend and improve the available toolkit for future VTT studies around the world. The recognition of the complex heteroskedasticity patterns, size and sign effects and deterministic heterogeneity has served to explain at least a share of the differences in valuations across respondents. Random (unexplained) heterogeneity of course remains, and the incorporation of attitudinal constructs and traveller specific constraints provide further directions for improvements.

Acknowledgements

The work reported in this paper is based on a study commissioned by the UK Department for Transport. The opinions and any omissions in this paper remain the responsibility of the authors. The authors are grateful for the comments made by the wider study team, including international peer reviewers, and also want to acknowledge the input of Michiel Bliemer to the experimental designs. The first author also acknowledges the support of the European Research Council through the consolidator grant 615596-DECISIONS for some of the additional analysis.

References

Abrantes P. and Wardman, M. (2011). Meta-analysis of UK values of travel time: An update. *Transportation Research Part A: Policy and Practice*. *Volume 45, Issue 1*, *p.1–17*

Accent Marketing and Research and Hague Consulting Group (1996). The Value of Time on UK Roads, *The Hague.*

Arup, ITS Leeds and Accent (2015) 'Provision of market research for value of time savings and reliability'. Phase 2 report to the Department for Transport. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/470231/vtts-phase-2-report-issue-august-2015.pdf

Association of Train Operating Companies (ATOC) (2012) 'Passenger Demand Forecasting Handbook'. Association of Train Operating Companies, London.

ATC (2006). National guidelines for transport system management in Australia [Online]. Available: http://www.atcouncil.gov.au/documents/

Austroad (2006). Update of RUC Unit Values to June 2005 [Online]. Available: https://www.onlinepublications.austroads.com.au/items/AP-T70-06

Batley, R., Ibañez, J.N (2013). Randomness in preference orderings, outcomes and attribute tastes: an application to journey time risk. The Journal of Choice Modelling 5(3), 157–175

Bliemer, M.C.J., Rose, J.M. and Chorus, C. (2014) Dominance in stated choice surveys and its impact on scale in discrete choice models. *Proceedings of the International Conference on Transport Survey Methods, Leura, Australia.*

de Borger, B. and Fosgerau, M. (2008) The trade-off between money and travel time: A test of the theory of reference-dependent preferences, J. of Urban Economics, Vol. 64, pp. 101-115.

Börjesson, M. and Eliasson, J. (2014). Experiences from the Swedish Value of Time study. *Transportation Research Part A, pp.144-158.*

Börjesson, M., Fosgerau, M. and Algers, S. (2012), Catching the tail: Empirical identification of the distribution of the value of travel time, Transportation Research. Part A: Policy and Practice, Vol. 46, No. 2, p. 378-391.

Carrion, C. and Levinson, D. (2012). Value of Travel Time Reliability: A Review of Current Evidence. *Transportation Research Part A 46, 720-741*

Daly, A. and Zachary, S. (1975) 'Commuters' values of time'. Report to Department of the Environment, LGORU report T55, January 1975; available at http://www.alogit.com/General_papers.htm.

Daly, A. (2010) 'Cost Damping in Travel Demand Models'. Report of a RAND Europe study for the Department for Transport, available at: http://webarchive.nationalarchives.gov.uk/20110202223908/http://www.dft.gov.uk/pgr/economics/rdg/costdamping/.

Daly, A., Tsang, F. and Rohr, C. (2014) The value of small time savings for non-business travel, in Transport Economics and Policy, Vol. 48, 2, May.

Department for Transport (DfT) (2013) 'TAG unit A1.3: User and provider impacts', November 2014. Available at: https://www.gov.uk/government/publications/webtag-tag-unit-a1-3-user-and-provider-impacts-november-2014

Engelson, L. and Fosgerau, M. (2016) The cost of travel time variability: Three measures with properties. *Transportation Research Part B*, 91, pp555-564.

Fosgerau, M. (2006). Investigating the distribution of the value of travel time savings, Transportation Research Part B, 40, pp. 688-707.

Fosgerau, M. (2007). Using Non-parametrics to specify a model to measure the value of travel time. *Transportation Research A 41 (9), pp.842-856*

Fosgerau, M. and Bierlaire, M. (2009) Discrete choice models with multiplicative error terms, Transportation Research Part B, 43, pp 494-505.

Fosgerau, M., Hjorth, K. and Lyk-Jensen, S. (2007a). The Danish Value of Time Study: Final Report.

Fosgerau, M., Hjorth, K. and Vincent, S. (2007b). An approach to the estimation of the distribution of marginal valuations from discrete choice data. *Munich Personal RePEc Archive, Paper No.3907*

Fosgerau, M. and Karlström, A. (2009) The value of reliability. *Transportation Research Part B*, 44 (1), pp38-49.

Fosgerau, M. and Mabit, S. (2013). Easy and flexible mixture distributions. Economics Letters, Volume 120, Issue 2, August 2013, Pages 206–210

Harris, A.J. and Tanner, J.C. (1974) . Transport demand models based on personal characteristics, Transport and Road Research Laboratory Supplementary Report SR65UC, Crowthorne, UK

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

Hess, S., Stathopoulos, A. & Daly, A.J. (2012), Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies, Transportation 39(3), pp. 565-591.

Hjorth, K. & Fosgerau, M. (2012), Using prospect theory to investigate the low marginal value of travel time for small time changes, Transportation Research Part B 46(8), pp. 917–932.

Hollander, Y. (2006). Direct versus indirect models for the effects of unreliability. Transportation Research Part A 40(9),699–711.

International Transport Forum (ITF) (2015) Valuing Convenience in Public Transport. Roundtable Report 156. OECD. Paris, France.

Kahneman, D. and Tversky, A. (1979) Prospect theory: an analysis of decision under risk. Econometrica, 47 (2), pp. 263-291.

Liu, X. and Polak, J.W. (2007) 'Nonlinearity and the specification of attitudes towards risk in discrete choice models', Transportation Research Record 2014, pp. 27-31

Mackie, P., Wardman, M., Fowkes, A.S., Whelan, G., Nellthorp, J. & Bates J.J. (2003). Values of Travel Time Savings in the UK. Report to Department for Transport. *Institute for Transport Studies, University of Leeds & John Bates Services, Leeds and Abingdon*. Available at: http://eprints.whiterose.ac.uk/2079/2/Value_of_travel_time_savings_in_the_UK_protected.pdf

Ojeda-Cabral, M., Batley R. and Hess, S. (2016). The value of travel time: random utility versus random valuation. Transportmetrica A: Transport Science, Vol.12, pp.230-248.

Ramjerdi, F., Flügel, S., Samstad, H., and Killi, M. (2010). Value of time, safety and environment in passenger transport – Time. *Transportøkonomisk institutt*

Rose, J.M. and Bliemer, M.C.J. (2014a) Stated choice experimental design theory: The who, the what and the why, in Hess, S. and Daly, A. (ed.) Handbook of Choice Modelling, Edward Elgar, Cheltenham, pp152-177.

Rose, J.M., and Bliemer, M.C.J. (2014) Survey artefacts in stated choice experiments. Proceedings of the International Conference on Transport Survey Methods, Leura, Australia.

Significance, VU and Bates (2013). Values of time and reliability in passenger and freight transport in the Netherlands. *Report for the Ministry of Infrastructure and the Environment.*

Train, K.E. (2009) Discrete Choice Methods with Simulation, second edition, Cambridge University Press, Cambridge, MA.

Tseng, Y.Y. and Verhoef, E.T. (2008) Value of time by time of day: a stated-preference study, Transportation Research Part B, 42 (7-8), pp. 607-618.