A critical appraisal of the use of simple time-money trade-offs for appraisal value of time measures

Stephane Hess
Institute for Transport Studies & Choice Modelling Centre
University of Leeds
s.hess@its.leeds.ac.uk

Andrew Daly
Institute for Transport Studies & Choice Modelling Centre
University of Leeds
andrew@alogit.com

Maria Börjesson
Centre for Transport Studies
KTH Stockholm
maria.borjesson@abe.kth.se

Abstract

Stated choice surveys have established themselves as the preferred approach for value of time elicitation with the help of choice models. However, major differences exist in the approach used across regions. While studies in Europe (particularly Northern Europe) continue to rely extensively on simple time-money trade-offs, those in Australia and South America in particular follow the notion that more complex and hence realistic setups are preferable. The present paper aims to provide a critical appraisal of the use of simple time-money trade-offs, drawing from our experience in recent European studies. We highlight a number of issues, in terms of differences in valuations across formats as well as a lack of clarity on how respondents actually interpret travel time in these simple time-money trade-offs.

Keywords: value of time; time-money trade-offs; stated choice

1. Introduction

Monetary valuations of travel time (VTT) form one of the main inputs into the appraisal of new transport infrastructure and policy schemes, and their robust estimation is therefore of considerable societal, financial and environmental importance. The need to derive robust estimates has also motivated important developments in the theory and practice of modelling choices and willingness to pay, developments that extend beyond the context of value of time to more general valuation.

While early studies (tracing back to Beesley, 1965 and Daly and Zachary, 1975) were based on real-world observation of travellers choosing between modes, the estimation of VTT has for the last three decades generally been performed via the analysis of Stated Choice (SC) data. At least in Northern Europe, and the UK in particular, the analysis of simple time-cost trade-offs has been prioritised. These countries (UK, Denmark, The Netherlands, Sweden & Norway) account between them for a very large share of national VTT studies. However, in other parts of the world, more complicated trading exercises have been preferred in national VTT studies, notably in Switzerland (Axhausen et al., 2008) and Germany (Ehreke et al., 2015). While an important argument in favour of simple trade-offs in applied work has been that other approaches are too complex/complicated, the counter-argument in academic work has been that what really matters is relevance (cf. Hensher, 2006). With the continuing reliance on simple trade-offs in some national studies, the issue therefore arises as to which approach should be adopted, especially if different results are obtained with different methods.

This paper discusses the theoretical and behavioural implications as well as empirical evidence and finds that there are indications that simple time-cost trade-offs give significantly different results and may be less satisfactory than more realistic exercises in obtaining the required values.

Before proceeding with our paper, it is worth briefly mentioning the history of these simple time-money trade-offs and provide some reasons for the resulting inertia in terms of methods. It was the national study of the value of travel time in the UK (MVA et al., 1987) that marked the acceptance of stated preference, including SC, for use in important transport policy work. The argument in the 1987 study was that SC gave acceptable error margins for VTT estimates with sample sizes that were feasible for the budgets available, while error margins from the revealed preference data available at that time were excessive. This study proved influential for example for the 1990 study for The Netherlands (Hague Consulting Group, 1990) and the following study for the UK (Accent and Hague Consulting Group, 1996). Perhaps unfortunately, a standard methodology was thus established and it then proved difficult for the civil services commissioning these important studies to deviate from the standard. For example, the subsequent Netherlands study (Hague Consulting Group, 1998) and the Swedish (Algers et al., 1995) and Norwegian (Ramjerdi, 1997) studies adopted very similar methods. In each case a simple binary time-money trade-off, sometimes presented as a route choice, formed a key component of the survey and analysis. Moreover, once repeated studies had been undertaken in a particular country, interest focussed on the change in VTT from the previous study, so that it was again very attractive for civil services to be able to compare results from a consistent methodology.

The above studies generally (though not always) also included scenarios other than simple time-money trade-offs, for example looking at the valuation of reliability or the valuation of travel time in different conditions of travel. However, the key appraisal values have been based on simple time-money trade-offs and results from other games (e.g. valuation of reliability) have been used simply as multipliers of the VTT measures from the time-money trade-offs, thus implicitly making the assumption that the base VTT is the same across formats.

As mentioned earlier, the use of simple time-money trade-offs in European national VTT studies differs substantially from the approach used elsewhere. Indeed, notably in Australian work (see e.g. Hensher & Rose, 2007), the argument has been put forward to present respondents with a single type of scenario (rather than different sets of choices across games) which includes all attributes that may be relevant in the choice, and to allow respondents to focus on what really matters to them (see e.g. Hensher, 2006). These studies generally also move away from a binary setting, typically relying on three alternatives. Including all attributes together in a single experiment also has the statistical benefit that the parameters for all attributes are estimated on the basis of the full data.

The argument of relevance of attributes presents a strong contrast to one of the key motivations for simple time-money trade-offs, namely that of reduced respondent burden, and producing estimates of VTT that are *unpolluted* by the presence of other attributes. In this context, two other pieces of empirical evidence are worth noting. Caussade et al. (2005) show that the variance of the error term in choice models seems indeed to increase with the number of attributes presented – this however simply means that there is more scope for other factors to influence choices, including heterogeneity in sensitivities for more attributes and/or differences in interpretation, and does not necessarily lead to biased estimates or different choice behaviour. More recent evidence (Chintakayala et al., 2010) suggests that increases in error only occur with more than 6 attributes. Evidence in Hess (2014) suggests little or no difference in willingness-to-pay (WTP) measures when additional attributes are included in the SC scenarios. The insights from these studies for the present paper are however limited – Caussade et al (2015) used different respondents for different treatments, while in Hess (2014), the full set of attributes was used for a respondent prior to focussing on a subset only, so directionality effects may exist. In typical European VTT studies, the simple experiment tends to be given first, and the directionality may thus be different. Much of this work is influenced by the study of Hensher (2004) who showed variations in WTP measures as a function of design dimensionality, while other work has shown how analysts can allow for the fact that some respondents may focus on a subset of the attributes only, i.e. filtering out information not relevant to them (e.g. Hensher, 2008).

Simple time-money trade-offs have often been the topic of heated discussions at conferences, with criticism notably from Australia and South America, but continue to be used in some national studies in Europe. In this paper, we revisit this issue on the basis of our own personal experience in the last 40 years, having contributed to a number of national studies that have relied on simple time-money trade-offs, as well as to more complicated valuation studies. This gives us a strong position to re-evaluate the reliability of this approach without simply attacking the work done by others. We revisit the pros and cons of the different approaches, and present new empirical evidence. In doing so, we also address an issue not discussed in the literature relying on simple time-money trade-offs, namely that of consistency of results across different formats of experiments (e.g. time-money vs time-money-reliability) presented to the same respondent. If differences arise across separate survey formats in the valuation for the same journey component, for example the VTT, then the question arises which of these should be used. Furthermore, if valuations for different journey characteristics (e.g. value of time vs value of reliability) are obtained from different survey components, and if any common valuations (typically the VTT) are not consistent across components, then this raises doubts about the wisdom of combining results from different components.

Our discussions focus solely on the setup of the choice tasks in terms of the attributes included (i.e. simply time-money vs more attributes) rather than the number of alternatives, which is a different discussion in itself. Similarly, we are not covering issues of experimental design, i.e. the specific time-cost combinations presented in a given choice task.

The remainder of this paper is organised as follows. Section 2 presents some a priori considerations in terms of advantages and disadvantages, while Section 3 explores the issues in more detail in the context of the recent British study. Section 4 presents additional evidence from other studies, before we summarise our findings in Section 5.

2. Advantages and disadvantages

We discuss the advantages and disadvantages of simple time-money trade-offs under three broad headings, namely a) micro-economic theory considerations, b) behavioural considerations, c) interpretation and d) methodological implications.

2.1. Micro-economic theory considerations

The value of travel time depends on three factors: the resource value of time (or opportunity value, the utility that could be attained if the travel time was used for some other activity), the direct utility of travel time (compared to some reference activity), and the marginal utility of money (DeSerpa, 1971; Jara-Díaz & Guevara, 2003; Jara-Díaz, 2007), with $\text{VTT} = \frac{\text{resource value of time} - \text{direct utility of travel time}}{\text{marginal utility of money}}$. The direct utility of travel time is influenced by factors such as the comfort and productivity or pleasure gained on the trip. The resource value of time increases the less available time the traveller has, and hence should vary depending on socio-economic status. The marginal utility of money is clearly related to the income.

The resource value of time depends on the traveller's time-use pattern and scheduling function, which may vary between trip contexts. A higher resource value of time is probably the main reason for higher value of time for commuting than other trip purposes. To capture the resource value of time the choice experiment is related to a reference trip that the respondent has recently undertaken. Respondents are instructed to respond to the choice experiment in the context of this reference trip and its related time-use pattern. We cannot disentangle the resource value of time from the direct utility of time. However, in all estimations, we model the variations in resource value of time across respondents by controlling for socio-economic variables such as trip purpose, travel time, family situation, employment, household size and obligations, and income, to name just a few. There are studies explicitly taking daily time scheduling adaptation process into account for variations in the resource value of time between and within people (Habib et al. 2013, Weis et al. 2013, Schmid, 2017)

Although the respondents are instructed to have the context of a specific trip in mind when making their choices, it is possible that the hypothetical setting implies that they pay less attention to the impacts of their choices on the time availability for other activities. Aside from investing the possibilities of revealed preference data, an interesting area for future work would be to test the differences in VTT that would arise from surveys looking not at a single trip but explicitly at the scheduling for an entire day. Even then, respondents may pay too little attention to the resource time of other days.

Assuming that the respondents follow the instructions and make the choices in the context of the reference trip, the VTT estimates produced from simple time-money trade-offs reflect the direct utility of travel time in the reference trip, including modal effects, but also reflect the resource value of time in this trip.[1] The differences in direct utility of time across modes and travel conditions can therefore only be explored by using an experiment in which the trip context varies for a given traveller and reference trip (for which the resource value of time stays constant).

Assuming a representative sample, the overall results can then be used as values of time (including the direct value of time and the resource value of time components) in current conditions. However, their use would prevent us from studying changes in the VTT as a result of changes in the direct utility that would arise from changes in travel conditions (e.g. a change in highway congestion or public transport crowding), as these conditions do not change across choice tasks or alternatives in simple time-money trade-offs.

2.2. Behavioural considerations

An initial key argument for simple time-money trade-offs has been that of reduced respondent burden, which was a common concern in the early days of SC surveys. Since then, important improvements have been made especially in the survey environment, for example because replacing paper-based surveys with web-based surveys allows a customized survey environment where for instance times can be calculated rather than showing +10 minutes etc[2]. Web-based surveys are also likely to reduce response error compared to telephone surveys (cf. Börjesson & Algers, 2011). More importantly, Chintakayala et al. (2010) suggest that there is no increased noise in data coming from more complex scenarios, and in fact suggest that overly simplistic scenarios may lead to behaviour that is more difficult to model (i.e. more randomness). The debate in academia about very complex surveys is ongoing, but here our focus is on VTT and the contrast between very simple and somewhat more complex games, and we thus do not revisit this issue in the present paper.

Strongly related to the point about respondent burden is the argument that in studies primarily (or solely) interested in estimating the VTT, the inclusion of other attributes in the experimental setup may simply act as a distraction, for both the modeller and the respondent. We return to the modelling implications later on, and for now focus solely on the behavioural effect. The argument that a simple time-money trade-off leads to an '*unpolluted*' measure of the VTT can quite reasonably be undermined by a counter-argument that this setup in fact leaves more space for interpretation open to the respondent, and hence an increase in noise. Indeed, respondents may enhance the information, for example inferring that a faster journey is less reliable (e.g. a faster route might be more popular and thus prone to accidents/congestion) or that a cheaper

[1] For instance, the differences in the VTT across modes arise from differences in the direct utility of time in the reference trip, due to differences in comfort and productivity of the travel time. However, the modal differences in the VTT arise also from self-selection, i.e. travellers with high resource value of time will tend to choose faster but more expensive modes.

[2] This is of course different from saying that web based surveys themselves have lower burden than paper based surveys, and the work on respondent burden (Axhausen et al., 2015) suggests this is indeed not the case.

train is less comfortable. Such unobservable (to the analyst) effects may lead to bias in the estimates that are of interest[3].

This point has directly led to the view by e.g. Hensher (2006) that what really matters is relevance. In other words, analysts should see possible increases in respondent burden, and/or the possibility of a respondent ignoring some of the presented attributes, as a lesser evil than the possibility of a respondent *enhancing* the presented information in an unobservable manner.

The focus on just time and money may also have two additional implications. Firstly, a key motivation for using hypothetical choice settings rather than transfer price/contingent valuation approaches is a recognition that the latter may lead respondents to purposefully bias their answers up or down depending on their vested interests. A simple time-money framework moves away from such direct elicitation, but is arguably still so transparent as to make it clear to respondents that they are being asked (albeit not directly) to give monetary valuations of travel time. Secondly, in real life contexts, respondents pay for improvements along a number of different dimensions (e.g. time, reliability, safety). To focus on just one in isolation can be imagined to either overstate the value (all the money can be spent on just travel time) or understate the value (an increase in cost is more visible in a two-attribute context).

Research in choice modelling has in recent years also focussed extensively on a number of behavioural effects, in particular non-linearity in preferences, and asymmetric preference formation around reference points (e.g. Hess et al., 2008; de Borger & Fosgerau, 2008), anchoring, lexicographic choice, inconsistent behaviour and non-trading (e.g. Hess et al., 2010) and heteroskedasticity (see discussions in Hess et al., 2016). These effects have been studied extensively in both simple and more complex trade-offs, but the question remains whether or not they are more prevalent in one or the other. This would lead to the conclusion that these effects are at least to some extent influenced by the experimental setup. A key question is whether respondents are actually not trading (i.e. they do not engage with the game) or whether the trade-offs presented to them are not large enough to incentivise them to move away from either the cheapest or fastest option. The former would of course raise issues as to the reliability of simple time-money trade-offs. It is in this context that the work of Börjesson et al. (2012) is highly relevant. They showed that in the Swedish study, extending the ranges of the trade-offs presented to respondents led to reductions in the rates of apparent non-trading, providing some reassurance that respondents do indeed engage with the format.

Our final point, and one that has attracted the most criticism of simple time-money trade-offs, is that of realism. In the introduction to this paper, we made the point that it is rare for a traveller to face a real-world choice between options that are distinguished only by one being a faster but more expensive option and one being a slower but cheaper option. While time-money trading may arise in a public transport context, it will often involve a mode choice (e.g. expensive but fast train vs cheap but slow bus) or different sub-modes (e.g. high speed vs conventional rail), thus bringing in implicit issues of comfort, reliability and status as well as the simple time and money differences. The core interest in many appraisal studies is on car, with the obvious real world context where a faster journey costs more arising in the case of toll

[3] Although less studied in the literature, it is conceivable that there is also an impact on the statistical accuracy of estimates, i.e. standard errors.

roads. To avoid strategic bias, most studies have however shied away from a toll road context (which is also not readily applicable in many countries anyway) and have attempted to address the realism issue by not talking about route choice but a choice between *different hypothetical contexts*. Whether a respondent can be realistically expected to make such a leap of faith is open to question.

The effects of the issues mentioned above are difficult to establish in empirical evidence given the latent nature of the behavioural processes that drive the results but also the fact that the data in different settings typically comes from different respondents. Our paper thus specifically focusses on comparisons where the same respondents provide data in more than one type of stated choice setting.

2.3. Interpretation

The first issue arising in the interpretation of results from simple time-money trade-offs is the question of what type of time is valued. There is a risk that if the travel context is not stated explicitly, it is not clear what conditions of travel the respondent has in mind when valuing the time, and this would in turn influence the direct utility of the value of time. In the interest of presenting as realistic a choice context as possible, increasing the chance of accurately capturing the resource value of time component of the VTT by making the respondents' time and money budget concrete, surveys often ask respondents to make choices as if they are in the context of a *current* or *recent* reference journey. The usual assumption is then that the VTT component reflecting the direct value of time coming out of time-money trade-offs reflects conditions similar or equal to those experienced on that reference journey. However, because the travel conditions are not explicitly stated it is far from certain whether this is the case. For example, a respondent may consider a change in time on the route as necessarily involving a change in congestion, and it is an assumption that has in our opinion not been questioned sufficiently thus far. It is on the other hand also possible that explicitly describing some of the travel conditions in the experiment gives rise to a focus effect related to the stated travel conditions. Empirical testing may not lead to conclusive findings, but we make some attempts later on in this paper.

The argument can be extended further to the cost attribute if one is willing to entertain the notion that respondents have different sensitivities to different cost components, as supported by much empirical evidence from surveys with multiple cost attributes (such as parking, toll and fuel costs for cars, for example) (see for instance Vrtic et al., (2007)). In passing we note that this would be problematic in itself since different sensitivities to different cost components are not micro-economically consistent.

Finally, if, as we hypothesised earlier, it is the case that respondents may *enhance* the information presented to them, by for example imagining that a faster journey is less reliable, then this is again likely to lead to issues with interpretation when focussing solely on changes in travel time and cost.

In the earlier section on behavioural effects, we have already alluded to the question whether phenomena such as reference dependence and non-linearity (commonly referred to as sign and size effects) are real world effects or survey artefacts. Namely, if we observe in empirical work based on stated choice data that a respondent reacts more strongly to increases in time than to decreases (compared to a reference point) and that the sensitivity to a larger change is smaller per unit of time than the sensitivity to a smaller change, then would the same hold for real world behaviour by that respondent?

Reference dependence is of course clearly also a real-world phenomenon, our question is simply whether what is picked up from data on hypothetical choices is the same as what happens in reality. A particular concern is that sign and size effects operate primarily in the short term, manifesting themselves only while there exists a well-defined reference point. In particular, value of time experiments implicitly require respondents to consider how to reschedule a reference trip to accommodate a travel time change in the short run, which is likely to increase reference dependence (Börjesson and Fosgerau, 2015). In the long run, e.g. the time horizon of a transport investment, there is however no stable reference point. Of course, in almost all cases, studies are required to produce a uniform VTT for appraisal that is independent of sign and size effects (cf. Daly et al., 2014), because in the long run there is no reference point. Moreover, long-term stability of preferences is a fundamental assumption in welfare economics. This means that findings of substantial sign and size effects are problematic. An important question is thus again whether sign and size effects are stronger or weaker in simpler or more complex trade-off contexts, a point we address in our comparisons later on.

Finally, as we mentioned in the introduction, a key argument behind the continued reliance on simple time-money trade-offs in many European studies has been that of continuity and comparability. This already hints at some underlying concerns that the values obtained with different experimental setups may vary (see also Widlert, 1994), and it is thus very surprising that only limited effort has gone into studying how the valuations obtained from the same respondents differ across context, something that is very different from studying differences in valuations across studies using different setups on different respondents. Our work described below addresses this issue.

2.4. Methodological implications

The methodological implications of the complexity of the experiments relate to both experimental design of the choice scenarios and the estimation of econometric models for the resulting data.

In terms of survey design, an initial argument for simple time-money trade-offs might have been a greater ability to ensure that the choices presented were all meaningful (i.e. avoiding complete dominance, where one option is both cheaper and faster, or quasi-dominance, where for example one option is marginally faster but extremely more expensive) and also covered the full spectrum of trade-offs (in terms of sign and size effects, i.e. looking at both increases and decreases in time and cost, and also changes of different amounts). While this argument would indeed have had some validity in the 1980s, experimental design techniques have evolved dramatically since, with algorithms and software that enable analysts to construct informative choice tasks with multiple attributes leading to data with good statistical properties.

The modelling implications fall into two broad categories, namely the detection of behavioural phenomena and the accommodation of them in the model. With both of these, the presence of just two attributes (and two alternatives) of course greatly simplifies the task faced by an analyst, and this can be seen as a key reason for the use of time-money trade-offs in a number of methodological studies in choice modelling.

With simple binary time-money trade-offs, a number of behavioural phenomena can be detected directly by *observing* the data. Examples of this include apparent lexicographic behaviour or non-trading (always choosing on the basis of just one attribute) and making inconsistent choices (e.g. accepting to pay for a faster journey in a time-money trade-off that implies a VTT of at least v_1 while later on refusing to pay in a trade-off that implies a VTT of not more than v_2, where $v_1 > v_2$). However, if sign and size effects are admitted, consistency of choices becomes complicated to analyse even in simple time-cost trading. The direct *inspection* of such issues becomes significantly more complicated in the case of more than two attributes (or more than two alternatives), and it is not generally possibly any longer to assign a single label to the observed behaviour. With simple binary time-money trade-offs, it also becomes possible to graphically inspect the data by plotting indifference curves (Fosgerau 2007, Börjesson et al. 2012) and this can then highlight the presence of heteroskedasticity and non-linearity/asymmetry in preferences. The ability to detect such phenomena directly in the data is of course an advantage that facilitates analysis and can help an analyst choose an appropriate model structure, but this should in no way mean that we no longer care about other possible disadvantages.

When it comes to the actual modelling of the behavioural phenomena, the presence of more than two attributes clearly complicates matters. This relates for example to the choice of reference points in terms of gain/loss asymmetry and the increase in the number of parameters to estimate. With a simple time-money trade-off, an analyst is also able to determine a boundary VTT in each choice task, and this can be used not only to study anchoring effects but can also be beneficial in models with more flexible error structures. However, once again, the study of these phenomena is also possible in the case of more than two attributes, as we highlight in the empirical section below. On the other hand, restricting the data to just time-money trade-offs (and in particular in a binary setting) also greatly limits the richness of the resulting data, given that there is less scope for variability and a lower dimensionality of the data. This can raise questions about the ability to reliably estimate models with distributed parameters.

3. New evidence from the 2014/2015 GB national VTT study

The recently completed national VTT study for Great Britain provides a useful testbed for the points discussed above. The study presented people using car, bus, rail or other public transport (e.g. tram or London underground) with different types of SC games. Full details of the experimental setup and the modelling work conducted are given in Arup et al. (2015), with detailed discussions also in Hess et al. (2017). For each of 11 purpose-mode combinations, multiple binary SC experiments involved different unlabelled trade-offs described by: time/money (SP1), time/money/reliability (SP2), and time/money/quality (SP3), where quality implied variation in either congestion or crowding. Respondents received all three games, with 5 choices per game, enabling us to make comparisons across games. In what follows, we focus on car and rail commuters, with samples of 922 and 847 respondents, respectively, although we also present some summary results for other mode-purpose combinations.

3.1. Behavioural effects

We first look at the rates of apparent time and cost non-traders across games, i.e. respondents who always choose the fastest (time) or cheapest (cost) option across the five choice tasks in a given experiment. We use the term 'apparent' as it is not clear whether a respondent really does not trade (and hence does not engage with the choice task) or whether their preferences are such that the incentive to move away from either the cheapest or fastest are not strong enough in the scenarios presented to them. Similarly, a respondent might in fact use a heuristic to never accept a cost above a certain amount or a journey taking longer than a given amount of time, and the scenarios presented might thus impose the choice.

Alongside game-specific rates, we also present the p-value for a χ^2 test for equality between the rates in the three games. Finally, we present the non-trading rates across the full set of 15 choices, which are then of course lower because they would have to be non-traders in each of the three games.

From the results in Table 1, an interesting picture emerges. We first observe that, across all four segments and across both time and cost non-trading, the χ^2 test rejects the null hypothesis of equality in the rates of non-trading across the three games.

For car commuters, we see that the rate of both time and cost non-trading clearly decreases in SP2 compared to SP1, and then again in SP3 compared to SP2. This is in line with having two attributes in SP1 (time & cost), three attributes in SP2 (time, cost and a measure of variability in time), and four attributes in SP3 (time across three conditions and cost).

The picture for other modes is slightly more complex. For time non-trading, we see that, except for other public transport (PT), the rate of non-trading is highest in SP1, while, except for other PT, SP2 and SP3 are similar. For cost non-trading, the rates in SP2 (which has more attributes) are lower than in SP1 across the three modes, but the SP3 rates are essentially the same as the SP1 rates. While this might seem counter-intuitive at first, it should be noted that SP3 here is a crowding game which presents only time and cost for each alternative, albeit that for time, a level of crowding is given. What we thus see is that the level of complexity of SP3 is very similar to that of SP1, and that the rate of those respondents who do not move away from the cheaper of the two options is not impacted by the additional information on crowding.

Overall, we find that Table 1 shows some evidence of higher rates of non-trading in simpler games, and this is confirmed by statistical tests.

Table 1: Non-trading in the 2014/2015 GB national VTT data

		Car commuters	Rail commuters	Other PT commuters	Bus commuters
Time non-traders (always fastest)	*SP1*	14.85%	10.19%	6.84%	8%
	SP2	12.82%	4.55%	7.84%	2.86%
	SP3[1]	10.87%	5.52%	0.53%	2.92%
	χ^2 *test p-value*	0.02	$2.7 \cdot 10^{-6}$	$1.1 \cdot 10^{-6}$	0.01
	over three experiments	1.26%	0.22%	0%	0%
Cost non-traders (always cheapest)	*SP1*	17.57%	19.16%	13.86%	22%
	SP2	11.06%	9.06%	5.17%	12%
	SP3[1]	8.34%	19.45%	12.44%	22.22%
	χ^2 *test p-value*	$7.6 \cdot 10^{-10}$	$1.2 \cdot 10^{-11}$	$1.5 \cdot 10^{-4}$	0.01
	over three experiments	2.04%	2.27%	1.06%	3.51%

1) crowding games only for rail, other PT and bus in SP3

A key emphasis in the modelling work conducted for the GB study was the modelling of sign and size effects, i.e. allowing for differences in how respondents react to increases and decreases from a reference value (say for time) and also in how the marginal sensitivity depends on the size of the shift presented. The specific approach used for this relies on the framework developed by de Borger and Fosgerau (2008), hereafter referred to as dBF, which provides all the flexibility we require.

In the dBF framework, we define a function that gives the value $v(\Delta x)$ of a change Δx relative to the reference value x_0 of a given attribute, as:

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

with $\Delta x = x - x_0$

where:

- $S(\Delta x)$ is the sign function, defined for $\Delta x \neq 0$ by $S(\Delta x) = \Delta x\,/\,(|\Delta x|)$; for convenience we set $S(0) = 0$.
- η captures sign effects by giving 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. It is expected that $\eta > 0$, so that the value of losses (increases in Δx) is greater than the value of gains.
- β captures size effect by allowing the impact of gains and losses to be non-linear, where, with $\beta > 0$, the marginal value of changes decreases as the change increases, i.e. the value is ‘damped’. 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 (size effect) of value to be different for gains and losses.

The 2014/2015 study allowed for sign and size differences in all three games, with the three parameters (η, β and γ) estimated for both time and cost. After extensive exploratory work (cf. Arup et al., 2015), the models were all specified with a multiplicative error structure, where for SP1, a log-WTP space specification was found to be preferable to a simple $U = V \cdot \varepsilon$

specification[4]. All models also allowed for random heterogeneity in the VTT across respondents.

Table 2 presents the estimates for the dBF parameters for car and rail commuters. The final specification of the models was arrived at via a long process which started with a fully specified model and gradually removed dBF terms that did not have a significant impact on the model. The decision on whether to remove a specific term was influenced not just by the statistical significance (i.e. t-ratio) but also the impact the removal had on the model fit and other parameters, and the size of the behavioural effect, i.e. whether for example the amount of non-linearity was 'behaviourally' significant as well as 'statistically' significant.

In the base specification, we would have estimated one β, one γ and one η term for both of time and cost and for each of the three games (i.e. 18 for each model). We note that:

- for size effects (β), half of the parameters are retained in SP1 as well as in SP2, while for SP3, only one out of four is;
- for asymmetrical size effects (γ), half of the parameters are retained in SP1, while only one in four are for SP2 and SP3; and
- for sign effects (η), i.e. differences between gains and losses, three out of four parameters are retained in SP1 and SP2, while only half are in SP3.

While the number of parameters in each model is not itself a statistical test, the key point here is that for each parameter, a statistical test was used to determine whether or not the parameter should remain in the model, and overall, this number is greater for SP1.

The presence of these additional terms complicates the VTT calculation, which is now no longer the simple ratio of time and cost coefficients. The value functions v (in equation 1) 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.

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 indifference value $\Delta c'$ such that $v(\Delta c') + v(\theta\Delta t') = 0$. The average willingness to pay per unit of time is then $\Delta c'/\Delta t'$.

As discussed in detail in Hess et al. (2017), the actual VTT is then given by

$$VTT = {}^{|\Delta c|}/_{|\Delta t|} = \theta^{\kappa}|\Delta t|^{\kappa-1}, \qquad (2)$$

where $\kappa = \frac{1-\beta_t}{1-\beta_c}$, and Δc is chosen such that $v(\Delta c) = -v(\theta\Delta t)$.

While η and γ present interesting behavioural insights, they drop out of the calculation of the VTT. The same is not true for the size effects represented by β. Only when $\beta_c = \beta_t$ are we in the situation where the VTT is independent of Δt, as the time and cost damping cancel out, i.e.

[4] We have no reason to believe that this difference explains the differences we report below, and the results obtained with $U = V \cdot \varepsilon$ on SP1 were broadly comparable, but with lower fit. Similarly, the differences we report across games were also confirmed by a simpler analysis using models with an additive error structure and without random heterogeneity.

we get that $\kappa = 1$. However, in general the β values will not be equal and VTT is not equal to θ but depends on the size of the time shift.

The most important impact of the dBF parameters in the VTT context is thus the fact that the VTT is no longer independent of the size of time shift considered if κ (given by $\frac{1-\beta_t}{1-\beta_c}$) is different from 1 in Equation (2). We see that, for car commuters, this is the case for SP1 and SP2, but more so for SP1 (i.e. the differences in VTT will be larger across different Δt in SP1 than in SP2), while there are no size effects for calculating VTT in SP3, which makes calculating appraisal values much easier. For rail commuters, we see values of κ different from 1 in all three games, where the size effects are slightly more important in SP2 than in SP1 and SP3.

Table 2: dBF parameters retained in final car and rail commuter models from 2014/2015 study

		Car commuters		Rail commuters	
		est.	**rob t-rat (0)**	**est.**	**rob t-rat (0)**
Size effects	$\beta_{t,SP1}$	-0.4000	-3.64	-0.2137	-3.31
	$\beta_{t,SP2}$	-0.1564	-2.84	-	
	$\beta_{t,SP3}$	-		-0.2418	-2.51
	$\beta_{c,SP1}$	-		--	
	$\beta_{c,SP2}$	-		0.2478	3.78
	$\beta_{c,SP3}$	-		-	
Asymmetric size effects	$\gamma_{t,SP1}$	-0.2127	-3.52	-0.1093	-2.39
	$\gamma_{t,SP2}$	-		-	
	$\gamma_{t,SP3}$	-		-0.1216	-2.00
	$\gamma_{c,SP1}$	-		-	
	$\gamma_{c,SP2}$	-		0.3456	3.40
	$\gamma_{c,SP3}$	-		-	
Sign effects	$\eta_{t,SP1}$	0.2573	4.34	0.0951	1.63
	$\eta_{t,SP2}$	0.0874	1.43	0.1408	2.18
	$\eta_{t,SP3}$	-		-	
	$\eta_{c,SP1}$	0.1267	2.18	-	
	$\eta_{c,SP2}$	-		0.2058	2.01
	$\eta_{c,SP3}$	0.2771	1.51	0.1379	1.90
coun		**est.**	**rob t-rat (1)**	**est.**	**rob t-rat (1)**
VTT size effects	κ_{SP1}	1.4	3.64	1.21	3.31
	κ_{SP2}	1.16	2.84	1.33	2.84
	κ_{SP3}	1	-	1.24	2.51

To add further evidence, Table 3 presents the values for β and κ across all mode-purpose combinations, excluding bus for which no size effects were observed, which can likely be linked to the fact that bus journeys are shorter. For other modes, we see that all 9 values for κ are different from 1 for SP1, while this reduces to 6 for SP2 and 3 for SP3. Additionally, for 6 out of the 9 cases where some of the values for κ are larger than 1, the highest values are obtained for SP1. Table 3 also reports significance levels for the differences in κ between the three different SP games for each segment. These calculations incorporate the correlations between the individual estimates. For comparisons where for one game, κ collapses to 1, the significance level for the difference (against 0) is of course the same as the significance level of a test against 1 for the κ which did not collapse to 1. We see that κ_{SP1} is significantly larger than κ_{SP2} at the 95% level or above in 3 out of the 9 segments, and just below the 90% level in one other segment. In comparison with κ_{SP3}, we see that κ_{SP1} is significantly larger at the 95% level or above in 4 out of the 9 cases, and around or above the 90% level in a further 2 cases. There is not a single segment where κ_{SP1} is significantly smaller than either κ_{SP2} or κ_{SP3}. Comparing κ_{SP2} and κ_{SP3}, we see that the former is significantly larger in three cases (where κ_{SP3} collapses to 1).

These results provide at least some suggestion that the incidence of size effects is greater for SP1 (i.e. simple time-money trade-offs) than for other games. It is also insightful to note that Arup et al. (2015) highlight differences across games in the incidence of significant estimates for η, i.e. showing the presence of asymmetries in the sensitivities to gains and losses. No effects were observed for bus. Looking at car, rail and other PT (tram, underground etc.), we note that, out of the 18 possible η parameters (time and cost three purposes and three modes), they found significant gain-loss asymmetry in 6 cases for SP1, 5 cases for SP2 and only 1 case for SP3.

Overall, these results provide some indication of differences across games in the role of reference formation, and may point to an influence of the experimental setup. Independently of whether these are real or experimental phenomena, the presence of these effects, especially size effects, poses significant issues in appraisal, and this could be seen as a disadvantage of simple time-money trade-offs such as SP1.

Table 3: β and κ values for all mode-purpose combinations for 2014/2015 GB study

	Car						Rail						Other PT					
	Commuters		**EB**		**Other non-work**		**Commuters**		**EB**		**Other non-work**		**Commute**		**EB**		**Other non-work**	
	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)
$\beta_{t,SP1}$	-0.40	-3.64	-0.11	-1.61	-0.14	-1.91	-0.21	-3.31	-0.15	-3.62	-0.13	-2.79	0		-0.15	-1.86	-0.12	-1.61
$\beta_{t,SP2}$	-0.16	-2.84	-0.45	-5.10	-0.24	-4.96	0		0		0		0		0		0	
$\beta_{t,SP3}$	0		0		0		-0.24	-2.51	-0.14	-3.14	-0.19	-3.73	0		0		0	
$\beta_{c,SP1}$	0		0.10	1.83	0.10	1.78	0		0.07	1.75	0		0.14	2.55	0.17	2.68	0	
$\beta_{c,SP2}$	0		0		0		0.25	3.78	0.14	1.95	0.16	2.84	0		0		0	
$\beta_{c,SP3}$	0		0		0		0		0		0		0		0		0	
	est	rob t (1)	est	rob t (1)	est	rob t (1)	est	rob t (1)	est	rob t (1)	est	rob t (1)	est	rob t (1)	est	rob t (1)	est	rob t (1)
κ_{SP1}	1.40	3.64	1.24	1.79	1.27	2.01	1.21	3.31	1.23	3.17	1.13	2.79	1.16	2.19	1.40	2.38	1.12	1.61
κ_{SP2}	1.16	2.84	1.45	5.10	1.24	4.96	1.33	2.84	1.16	1.69	1.19	2.39	1	-	1	-	1	-
κ_{SP3}	1	-	1	-	1	-	1.24	2.51	1.14	3.14	1.19	3.73	1	-	1	-	1	-
	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)	est	rob t (0)
κ_{SP1}-κ_{SP2}	0.24	2.47	-0.21	-1.39	0.02	0.17	-0.12	-0.77	0.07	0.64	-0.06	-0.52	0.16	2.19	0.40	2.38	0.12	1.61
κ_{SP1}-κ_{SP3}	0.40	3.64	0.24	1.79	0.27	2.01	-0.03	-0.27	0.09	1.04	-0.05	-0.92	0.16	2.19	0.40	2.38	0.12	1.61
κ_{SP2}-κ_{SP3}	0.16	2.84	0.45	5.10	0.24	4.96	0.09	0.54	0.01	0.13	0.00	0.03	-	-	-	-	-	-

3.2. Consistency across survey contexts

The other key issue raised in the earlier parts of this paper was that of interpreting what the time attribute in simple time-money trade-offs actually represents, i.e. what types of conditions respondents associate it with. A simple interpretation could be that the type of time being valued is representative of the conditions experienced by the respondent on the reference journey, i.e. if this journey was in heavy congestion, then the direct utility of the time component of the VTT from a simple time-money trade-off would relate to time spent in heavy congestion. The 2014/2015 study attempted to explain differences across different respondents in their valuations by linking them to conditions on the reference journey. However, to do this it is necessary to assume that the resource value of time is identical for respondents experiencing different conditions on the reference journey, since we cannot explicitly disentangle the resource value of time from the direct utility of time.

For example, for rail travellers, the study tested whether respondents who had reported high levels of crowding for their reference journey had higher VTT in SP1 than those who had reported low levels of crowding. Similar tests were made for car travellers, trying to relate the SP1 valuations to the level of congestion experienced on their reference trip. Overall, these attempts were not successful[5]. This can of course simply be interpreted as self-selection, i.e. respondents who travel on crowded trains are less sensitive to crowding than other respondents, with a similar argument for those people who drive during congested times. It can also be interpreted as saying that respondents who travel on crowded trains have a lower resource value of time (although that does not seem likely). However, the finding for SP1 means that we do not learn anything about how valuations differ across different travel conditions from simple time-money trade-offs. Such information can only be obtained by understanding how the same respondent, considering the same reference trip (so that the resource value of time stays constant), reacts to different journey conditions. Taking a value estimated from a simple time-money trade-off may of course tell us the average VTT across travellers in current journey conditions (providing the sample is representative) but we cannot learn anything about how that behaviour might change if journey conditions are changed.

We next compare the core VTT measures across the three games for each mode, focussing on commuters only. The results for this comparison are shown in separate panels for the four modes in Figure 1. Noting that values for κ different from 1 arise for car, rail and other PT, the VTT measures are shown for different values of Δt (cf. Equation (2)). This complication does not arise for bus.

Looking first at car, we note the absence of size effects for SP3, and hence the flat profiles for the value of time in the three conditions. We also include a value weighted according to average conditions in the UK. With κ_{SP1}=1.4 and κ_{SP2}=1.16, we note increasing VTT for SP1 and SP2 (but less so for the latter) with increasing Δt, underlining the difficulty of using these values for appraisal. More importantly, however, for low values of Δt, the SP1 value is below that for free flow conditions, while with high Δt, it is close to that in heavy congestion.

[5] Including in further tests reported in http://www.stephanehess.me.uk/papers/Hess_impact_of_congestion_on_SP_values.pdf

As discussed at length in Arup et al. (2015), the presence of these size effects and their impact on the VTT leads to the difficult situation in which the analyst needs to make a choice for a value of Δt. This has impacts on the relative values across modes and is thus not a decision to be taken lightly. It also enhances our view that choice contexts that do not lead to significant size effects greatly simply the use of results. With the value of $\Delta t = 10$ chosen in the UK appraisal framework, the VTT from SP1 is above the SP3 values weighted to average conditions, while that from SP2 is close to that for high congestion in SP3. This highlights the earlier points about the difficulty faced in interpreting values from simple time-money trade-offs, but also extends it to time-money-reliability trade-offs where the type of time being valued is not explicitly described.

Looking next at rail, we have relatively similar values for κ across the three games, and thus similar non-linearities. We present the simple VTT from SP1 and SP2, along with the VTT in the lowest and highest crowding conditions (out of 10 conditions) from SP3, and a value weighted to average rail crowding in the UK. We note that the value from SP1 is again below that for SP2 (except for very low Δt), while it is also below that for average crowding conditions.

For bus commuters, no size effects were observed and we thus present a single value for each component. We again observe a higher value for SP2 than for SP1, and note that the value for SP1 is below the value for free flow time in SP3, and below the values weighted to either traffic or crowding conditions (noting that some bus users received a crowding game as SP3 while others received a congestion game as SP3).

Finally, for other PT, size effects are only observed in SP1, and this shows that the value for SP1 rises with Δt, going from below that in the lowest crowding conditions, to exceeding average conditions with a value of Δt of below 5, while, at the value of $\Delta t = 10$ used in appraisal, the value from SP1 equals that from SP2 and exceeds that in average conditions.

Across the four modes (and similar results arise with other purposes), we can reach a number of conclusions:

- the VTT from SP2 is generally higher than that from SP1, suggesting that, in the presence of reliability information, respondents place a lower sensitivity on cost than in the absence of reliability information;
- there is no evidence to suggest that the VTT from SP1 corresponds in any way to the values from SP3 weighted to average conditions, and in fact the former is generally lower than the latter; and
- the presence of size effects significantly complicates the picture, leading to further issues in choosing values for appraisal, and differences across games in the relationship between values across different games.

Overall, this discussion adds empirical evidence to our earlier concerns about what the VTT from SP1, i.e. the simple time-money trade-off, “means”.

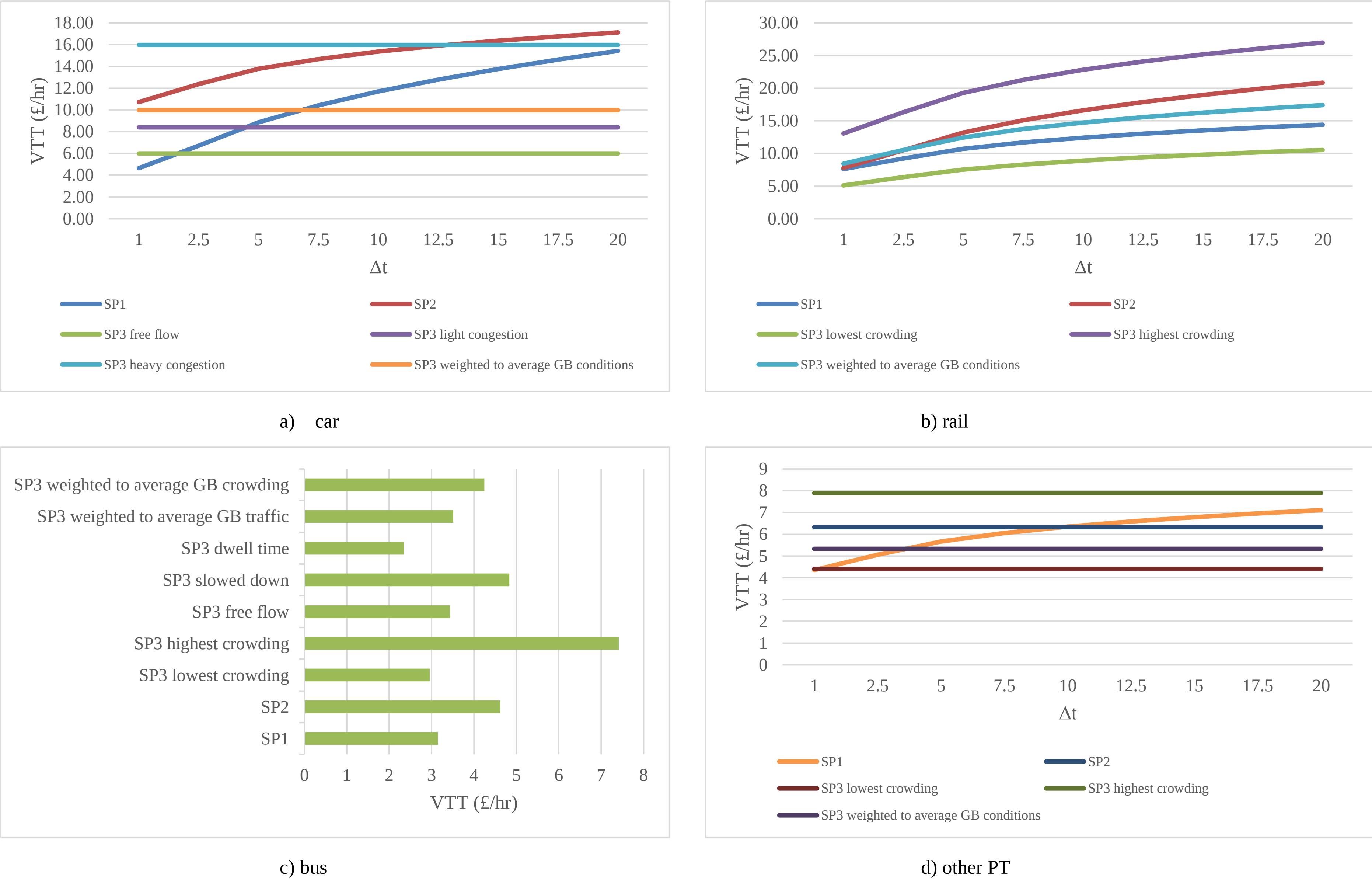

Figure 1: comparison of VTT across games for commuters in 2014/2015 GB study

4. Comparison with evidence from other studies

To provide some broader evidence, we now also briefly summarise results from other national VTT studies, commissioned by the central governments of the countries concerned, and using similar survey approaches to those described above. In particular, they all included a simple time-money trade-off, and the recommended values are always based on that trade-off, with the exception of the Netherlands study, as discussed below.

4.1. UK study (1994 data)

The previous UK study, based on 1994 data, was described by Accent and Hague Consulting Group (AHCG, 1996) and a re-analysis of the data by ITS and Bates (2003). The re-analysis used greatly simplified models, based on part of the data only, compared with the original study and we shall base our discussion primarily on AHCG.

Up to three experiments were presented in the study, which considered road users only. Of interest for the present paper are the simple time-money trade-off and a separate experiment offering a choice between tolled and untolled routes with a time difference and a cost difference partially due to the toll.

VTT was found to vary with the amount of time difference offered to respondents, as in many other studies. After consideration, small time differences were not included in the final calculation. While the analysis relied on models without random heterogeneity, extensive interactions with covariates were included

The headline values from the two experiments discussed above are shown in Table 4, alongside confidence limits obtained by a Jack-knife procedure.

The results suggest similar valuations for commuters in the two experiments, while, for business and other, the valuations differ between the two games, but with differences in opposite directions for the two purposes. No conclusive picture thus emerges, other than that differences may exist.

Table 4: AHCG values of car driver travel time (mean values, 1994 p./min.)

	Commuter	Business	Other
VTT from time-money trade-off	5.4 ± 0.9	11.9 ± 1.1	4.4 ± 0.7
VTT from toll road experiment	5.4 ± 1.4	8.0 ± 1.7	5.8 ± 1.6

Note: the error margins are for 95% confidence.

4.2. Danish study (2004 data)

In this study, data was collected in four stated preference exercises which related to a specific journey made by the respondent (Fosgerau et al., 2007). We again compare the results for a simple time-money trade-off to those from a game involving both in-vehicle and out-of-vehicle components. For car drivers, the components were walking time to/from a car park, searching time for a parking space, free-flow driving and congested driving time, while, for public transport users, the experiment covered in-vehicle time, headway, access/egress time, interchange times as well as cost.

The study introduced the concept of working in log-WTP space for the time-money trade-offs and a multiplicative specification for the more complex game, as also done in the recent GB study. The dBF approach was used to allow for sign and size effects in the time-money trade-offs. However, reference effects were limited to the influence of sign, i.e. no asymmetric size effects. In both experiments, the central VTT value was assumed to be randomly distributed in the population. As in the GB study, the presence of size effects meant that a decision on Δt was required, and a value of 10 minutes was used, which is of course again arbitrary.

Table 5: DATIV values of travel time
mean values, all purposes (not working) 2003 DKK per hour, distributions truncated at DKK 1000/hr

	Car driver free-flow	Car pass-enger free-flow	Bus	Metro	S-Train	Train
VTT from time-money trade-offs	78	52	30	62	35	54
VTT from more complex trade-offs	<50 km: 98 ≥50 km: 78	<50 km: 78 ≥50 km: 98	37	60	54	<50 km: 52 ≥50 km: 183

Note: error margins are not given in the published report.

Although the pattern of VTT in the more complex trade-offs for car driver is unusual, as longer journeys usually have a higher VTT, and the Metro values are contrary to the general trend, it is clear that the more complex trade-offs generally yield a somewhat higher VTT than the simple time-money trade-offs. Since error margins are not quoted in the reports, we are unable to say whether this effect is statistically significant.

4.3. Netherlands study (2009 and 2011 data)

This study is described by Significance et al. (2013). Two experiments dealt with car driver VTT. The first was the usual time-cost trading exercise, while the second involved time, cost and variation in travel time, with some variation in arrival time in part of the experiment. The two experiments were analysed together, assuming that the VTT is equal across the experiments. We are informed (private communication) that the values were not found to be significantly different across the games in early analyses, which however relied on simpler models than the final structures. It remains open to further analysis whether equality of values across games also applies in the more advanced models, but this is far from certain given that the values from advanced models often differ substantially from more basic ones.

4.4 The Norwegian Study (2009 data)

In the Norwegian 2009 VTT study (Ramjerdi et al. 2010), each respondent received 3 choice experiments. First all respondents received 9 choice tasks from a simple time-money trade-off, followed by two sets of 6 choices involving an additional attribute related to either quality or reliability. The values are summarised in Table 6 and, although the models used differed between the experiments, there is again evidence of different values of time in the simple time-money trade-offs.

Table 6: Values of time (mean values, 2009 NOK per hour)

		Car	Rail	Bus	Air	PT
VTT from time-money trade-offs	Long-distance	148	97	75	192	
	Short-distance	84				54
VTT from congestion game	Long-distance	157				
	Short-distance	47				
VTT from seat availability game	Long-distance					
	Short-distance					36
VTT from mean variance game	Long-distance	255	268	150	522	
	Short-distance	128				67
VTT from scheduling model game	Long-distance	95	100	223	235	
	Short-distance	35				31

Note: error margins are not given in the study report.

5. CONCLUSIONS

This paper has sought to provide a critical assessment of the continued reliance on simple time-money trade-offs in many national value of time studies in Europe, when such approaches are routinely criticised elsewhere as being too abstract and potentially prone to bias.

The aim behind this paper was in no way to invalidate the large amount of high quality methodological work in choice modelling that has made use of such data in recent years, but rather to question its reliability for producing values for appraisal. In this context, we specifically ask the question whether the fact that such approaches have been used many times in the past is a valid justification for continuing to rely on them.

We have highlighted issues in terms of realism in simple time-money trade-offs especially for car travel (a faster route would generally be cheaper) as well as concerns about interpretation of such trade-offs by respondents in the absence of a clear definition of the type of time being valued. Empirical support for this latter point was provided by the results from the recent GB study.

Our overview of the results from a number of recent studies has provided further evidence to support the notion that the VTT estimates differ across formats used in the SC surveys. We can conclude that the VTT from simple time-money trade-offs tend to be lower (sometimes substantially lower) than those from other settings, maybe suggesting that with just two attributes, respondents focus more on cost. This would also be consistent with the findings in Hensher (2004). But in some of the studies, the value from the simple time-money trade-offs is higher for some of the modes.

Of course, we do not know what value is *"right"*, and one could of course argue that rather than the value from simple time-money trade-offs being too low, that from more complex surveys is too high (or vice versa). We do not have an a priori truth and our decision on which value should be used needs to be guided by which one we trust more and which seems more reliable. In this context, a number of other considerations should be taken into account.

Firstly, it is clear that with simple time-money trade-offs, we risk producing a VTT measure that relates to travel conditions which we do not know and that the variation in that value across respondents does not reflect the variation in conditions across respondents. Indeed, even if we can link them back to experienced real-world conditions, issues with self-selection remain, and we learn nothing about how the same person would react to different conditions. This creates issues in terms of recognising differences in the direct utility of travel time (cf. Section 2.1)[6].

Secondly, the results from the most recent UK study also suggest that there is increased scope for reference dependence and non-linearity in data from simple time-money trade-offs. While these effects are behaviourally interesting, they lead to great complications in using the outputs in appraisal. Even if they are *real-world* effects arising from the framing of the experiment around a short-term setting, and not artefacts from the hypothetical setting, they are a severe problem since welfare analysis relies on long-term stable preferences.

Finally, an important point we touched on in the introduction is that appraisal is of course interested not just in the monetary valuation of travel time but also numerous other journey components. If the VTT is obtained from simple time-money trade-offs and then combined with other measures (e.g. value of crowding) from more complex settings, then there is a significant risk that the different values are not compatible with each other. As an example, assume that the VTT from a simple time-money trade-off is £10/hr, while the VTT from a game also incorporating interchanges is £15/hr and shows that one interchange is valued in the same way as 15 additional minutes in time. If the appraisal framework then uses the VTT of £10/hr, should interchanges be valued at £2.50 or £3.75? Significant issues with fungibility may arise (cf. Hess et al., 2012) and there might be benefits by instead combining all components into one survey.

[6] In other evidence for this, it is worth noting that the 2008 Swedish VTT study used a method resembling the one used by Ortúzar (2007) to value the perceived security when accessing PT. Respondents were presented with two binary stated choice experiments, differing in the dimensions of in-vehicle travel time, walk time and wait time. In the first experiment, the respondents were asked to have a recently made PT trip in mind. In the second experiment, one of four different physical walking environments was attached to each choice task, presented by coloured drawings. The walking time weights were found to be consistently higher in the second experiment than in the first, even for the nicest and most secure walking environment. This could be due to an over-focus effect on the walking time in the second experiment. However, it is also a support of the conclusion from the UK study, indicating that we cannot learn about the how the value of time depends on travel conditions unless we explicitly vary them in the experiments.

Overall, in the face of recent academic work suggesting that a) respondents can deal with complex choice scenarios, b) the valuations from more complex and therefore more realistic scenarios are reliable and c) that analysts can accommodate differences in how respondents relate to these more complex environments, our suggestion would be for applied work used to guide national policy to move towards more complex and realistic games.

ACKNOWLEDGEMENTS

The first author acknowledges the financial support by the European Research Council through the consolidator grant 615596-DECISIONS. The last author acknowledges the financial support from VINNOVA grant 2014-03587.

CONTRIBUTIONS

Stephane Hess: literature review, empirical work, manuscript preparation and editing

Andrew Daly: literature review, manuscript preparation and editing

Maria Börjesson: literature review, manuscript preparation and editing

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Accent and Hague Consulting Group (1996): The value of travel time on UK roads, London, Report for Department of Environment, Transport and the Regions.

Algers, S., J. Lindqvist Dillén, and S. Widlert (1995): 'The national Swedish value of time study', Presented to 'Traffic Days', Aalborg University.

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

Axhausen, K.W., Hess, S., König, A., Abay, G., Bates, J.J. & Bierlaire, M. (2008), Income and distance elasticities of values of travel time savings: New Swiss results, Transport Policy, 15(3), pp. 173-185.

Axhausen, K. W., Schmid, B., Weis, C., 2015. Predicting response rates updated. Arbeitsberichte Verkehrs-und Raumplanung 1063.

Beesley, M. E. (1965): 'The value of time spent in travelling: some new evidence', Economica, 32, 174–85.

Borger, B. de 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. & Algers, S. (2011), Internet actually reduced response error: Properties of internet and telephone data collection methods in a stated choice value of time study context, Journal of Choice Modelling, 4(2), pp. 1-19.

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, pp. 378-391, 2012.

M. Börjesson, "Valuing perceived insecurity associated with use of and access to public transport," Transport Policy, vol. 22, pp. 1-10, 2012.

Börjesson, M. and Fosgerau, M. "Response time patterns in a stated choice experiment," Journal of Choice Modelling, vol. 14, pp. 48-58, 2015.

Burge, P., Rohr, C., Vuk, G. and Bates, J. (2004) Review of international experience in VOT study design, presented to European Transport Conference.

Caussade, S., Ortúzar, J. de D., Rizzi, L.I. and Hensher, D.A. (2005) "Assessing the influence of design dimensions on stated choice experiment estimates". Transportation Research B, 39, pp. 621-640.

Chintakayala, P.K, Hess, S., Rose, J.M. and Wardman, M.R. (2010), Effects of Stated Choice Design Dimensions on Model Estimates, in Hess, S. & Daly, A., Choice Modelling: the state-of-the-art and the state-of-practice; Proceedings from the Inaugural International Choice Modelling Conference, Emerald, Bingley, pp. 195-216.

Daly, A. J. and Zachary, S. (1975) *Commuters' Values of Time*, Local Government OR Unit Report T55, January; available at http://www.alogit.com/papers/CommutersValues.pdf, accessed 29 July 2016.

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.

DeSerpa, A. C. (1971). A Theory of the Economics of Time. The Economic Journal, 81(324), 828–846. https://doi.org/10.2307/2230320

Ehreke, I., Hess, S., Weis, C. & Axhausen, K.W. (2015), Reliability in the German value of time study, Transportation Research Record, 2495, pp. 14-22.

Fosgerau, M., Hjorth, K. and Lyk-Jensen, S. (2007) The Danish Value of Time Study, DTU Report 5, Final Report; Note 5, Results for Experiment 1; Note 6, Results for Experiment 2; all available at
http://www.dtu.dk/english/Service/Phonebook/Person?fr=101&id=38996&mr=100&qt=dtupublicationquery&tab=2, accessed 30 July 2016.

Habib, Khandker Nurul, Ana Sasic, Claude Weis and Kay W. Axhausen. 2013. Investigating the nonlinear relationship between transportation system performance and daily activity–travel scheduling behaviour. *Transportation research. Part A, General* 49: 342-357.

Hague Consulting Group (1990): The Netherlands Value of Time Study: Final Report, The Hague: Dienst Verkeerskunde, Rijkswaterstaat.

Hague Consulting Group (1998): Value of Dutch Travel Time Savings in 1997: Final Report, The Hague, Adviesdienst Verkeer en Vervoer, Rijkswaterstaat, HCG Report 6098.

Hensher, D.A. (2004) Accounting for Stated Choice Design Dimensionality in Willingness to Pay for Travel Time Savings, Journal of Transport Economics and Policy, 38 (2), 425-446.

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

Hensher, D.A. and Rose, J. (2007) Development of Commuter and non-Commuter Mode Choice Models for the Assessment of New Public Transport Infrastructure Projects: A Case Study, Transportation Research A, 41 (5), 428-433

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

Hess, S., Rose, J.M. & Hensher, D.A. (2008), Asymmetric Preference Formation in Willingness to Pay Estimates in Discrete Choice Models, Transportation Research Part E, 44(5), pp. 847-863.

Hess, S., Rose, J.M. & 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. (2014), Impact of unimportant attributes in stated choice surveys, European Journal of Transport and Infrastructure Research, 14(4), pp.349-361.

Hess, S., Orr, S. & Sheldon, R. (2012), Consistency and Fungibility of Monetary Valuations in Transport, Transportation Research Part A 46(10), pp.1507-1516.

Hess, S., Daly, A., Dekker, T., Ojeda Cabral, M. & Batley, R.P. (2017), A framework for capturing heterogeneity, heteroskedasticity, non-linearity, reference dependence and design artefacts in value of time research, Transportation Research Part B, 96, pp. 126–149.

Jara-Díaz, S. (2007) Transport economic theory. Emerald Group Publishing Limited.

Jara-Díaz, S. R., & Guevara, C. A. (2003). Behind the Subjective Value of Travel Time Savings. Journal of Transport Economics and Policy, 37, 29–46.

MVA, ITS Leeds, and TSU Oxford (1987): 'The value of travel time savings', Policy Journals, Newbury, Berkshire.

Ortúzar, J.D. (2007): Valuation of intangibles with stated preference methods: A view from the fringe of the empire. Oslo Workshop on Valuation Methods in Transport Planning, March 19-20, 2007.

Ramjerdi (1997): The Norwegian Value of Time Study (online).

Ramjerdi, F., Flügel, S., Samstad, H., & Killi, M. (2010). Value of time, safety and environment in passenger transport ? Time - Transportøkonomisk institutt. Retrieved from http://www.toi.no/article29726-29.html

Schmid, Basil, and Kay W. Axhausen. Investigating suppressed demand effects in a Post Car World: A latent variable random effects Poisson (LVREP) approach. Paper presented at the *17th Swiss Transport Research Conference (STRC 2017)*, Ascona, Switzerland, May 17-19, 2017

Significance, VU Amsterdam and Bates, J. (2013) Values of time and reliability in passenger and freight transport in The Netherlands, report for Ministry of Infrastructure and the Environment.Vrtic, Milenko, Nadine Schüssler, Alexander Erath and Kay W. Axhausen. 2007. Route, mode and departure time choice behaviour in the presence of mobility pricing. Paper presented at the 86th Annual Meeting of the Transportation Research Board, January 21–25

Weis, Claude, Christoph Dobler and Kay W. Axhausen. 2013. A stated adaptation approach to surveying activity scheduling decisions. In Transport survey methods : best practice for decision making, Johanna Zmud, 569-589. Bingley: Emerald, November 14-18.

Widlert, S. (1994), Stated Preference Studies – The Design Affects the Results, 7th IATBR conference, Chile.