Not bored yet – revisiting respondent fatigue in stated choice experiments

Stephane Hess
Institute for Transport Studies
The University of Leeds
Leeds UK
S.Hess@its.leeds.ac.uk

David A. Hensher
Institute of Transport and Logistics Studies
The Business School
University of Sydney
NSW 2006 Australia
David.Hensher@sydney.edu.au

Andrew Daly
Institute for Transport Studies
The University of Leeds
Leeds UK
RAND Europe, Cambridge, UK
daly@rand.org

Abstract

Stated choice surveys are used extensively in the study of choice behaviour across many different areas of research, notably in transport. One of their main characteristics in comparison with most types of revealed preference (RP) surveys is the ability to capture behaviour by the same respondent under varying choice scenarios. While this ability to capture multiple choices is generally seen as an advantage, there is a certain amount of unease about survey length. The precise definition about what constitutes a large number of choice tasks however varies across disciplines, and it is not uncommon to see surveys with up to twenty tasks per respondent in some areas. The argument against this practice has always been one of reducing respondent engagement, which could be interpreted as a result of fatigue or boredom, with frequent reference to the findings of Bradley & Daly (1994) who showed a significant drop in utility scale, i.e. an increase in error, as a respondent moved from one choice experiment to the next, an effect they related to respondent fatigue. While the work by Bradley & Daly has become a standard reference in this context, it should be recognised that not only was the fatigue part of the work based on a single dataset, but the state-of-the-art and the state-of-practice in stated choice survey design and implementation has moved on significantly since their study. In this paper, we review other literature and present a more comprehensive study investigating evidence of respondent fatigue across a larger number of different surveys. Using a comprehensive testing framework employing both Logit and mixed Logit structures, we provide strong evidence that the concerns about fatigue in the literature are possibly overstated, with no clear decreasing trend in scale across choice tasks in any of our studies. For the data sets tested, we find that accommodating any scale heterogeneity has little or no impact on substantive model results, that the role of constants generally decreases as the survey progresses, and that there is evidence of significant attribute level (as opposed to scale) heterogeneity across choice tasks.

Keywords: fatigue; stated choice experiments; multiple data sets; willingness to pay; scale; learning

1. Introduction

Choice experiments are a popular setting within which to investigate the preferences of a sample of individuals between offered alternatives, where one or more alternatives may be available today and/or are prospective offers. These surveys are used extensively in producing guidance for policy makers and thus play a crucial role in transport policy and planning. There is an extensive literature on the design of choice experiments (see Bliemer & Rose, 2009, for a detailed review), and a growing literature on investigating candidate attribute processing rules invoked by respondents (see Hensher, 2010, for a detailed review) when asked to evaluate a series of alternatives in a given choice set and to choose the most preferred (or to rank all the alternatives). The repeat nature of the choice experiment across a number of choice sets has been recognised as a feature that requires special attention, with the focus primarily on ways of accounting for the correlated structure induced by offering each respondent multiple choice sets in a sequence (see Daly & Hess, 2010, for a recent discussion).

In addition, there has been great interest in the possible role of three behavioural mechanisms as a respondent goes through a survey, namely fatigue, boredom, and learning. The impact of these mechanisms on results is typically thought to be restricted to model scale. If a respondent is bored or becomes fatigued, his or her engagement with the survey reduces or mistakes are being made, and as a result, model scale goes down (i.e. the variance of the error increases). Fatigue and boredom, though different in nature, thus have a possibly very similar impact on results, and in practice it will not be possible to distinguish between them. In the remainder of the paper, we will largely be restricting ourselves to the term 'fatigue'. Learning on the other hand would mean, most simply, that as a respondent starts to better understand the choice tasks, model scale increases. Fatigue and boredom especially have received a lot of attention, but the claims and the suggestion that the error variance associated with the alternatives defining a choice set increases throughout the sequence (or in some studies that it is U-shaped), are typically based on a single data set. In this context, a disproportionately large weight is given, especially in applied studies, to the early evidence in Bradley & Daly (1994). This paper showed that in a survey making use of up to sixteen binary choice tasks (thirteen on average), the scale decreased significantly throughout the duration of the experiment.

The issue of fatigue has also been addressed by a number of other authors. Adamowicz, et al. (1998) looked at the issue in a study involving eight choice sets with three alternatives each, and found no effects of either learning or fatigue. Phillips et al. (2002) use a 12-task experiment and find lower scale in the second set of six tasks than in the first set of six tasks. Hanley et al (2002) found no significant differences between the results for respondents who faced four tasks and the results for respondents who faced eight tasks in a five-attribute, two-alternative experiment. Risa Hole (2004) also made use of choice set specific scale parameters in a data set collected for forecasting the demand for an employee Park and Ride service, where each respondent was shown nine choice sets each containing only two attributes, time and cost. None of the scale parameters were significantly different from one. Holmes and Boyle (2005) look at an experiment with four choice tasks and find higher scale for the fourth task than for the first task.

Savage and Waldman (2008) investigated learning and fatigue across multiple stated choice (SC) tasks for both mail and online survey modes, and found that out of the two, only online respondents responded with decreasing quality as they progressed through multiple choice tasks. Caussade et al. (2005) investigated effects of learning and fatigue while simultaneously investigating the impact of design complexity and cognitive burden in a heteroskedastic logit framework on a route choice data. It was observed that learning effects were prevalent during the first nine tasks, followed by fatigue effects. The notion of learning effects being followed by fatigue effects is also consistent with the findings by Hu (2006) who interacted the scale parameter with the task number in a mixed logit

model and found learning effects over the first six choice tasks, followed by fatigue effects. Brazell and Louviere (1996) similarly suggested that scale increases up to a point before it decreases again, in an experiment with 96 tasks, while Raffaelli et al (2009) found that learning effects dominated over the first ten out of sixteen choice tasks in a five attribute five alternative experiment making use of the heteroskedastic logit model. Bateman et al. (2008) use a sixteen-choice-task experiment and find that only the scale parameters for tasks 3, 9, and 16 are significantly different from 1, where contrary to common assumptions, the least reliable estimates are in the middle of the sequence, with the most reliable question being the sixteenth one, a point by which most analysts would assume fatigue to have set in. Bech et al. (2010) exposed respondents to different numbers of choice tasks (5, 9, 13, and 17) and found that respondents who were given 17 tasks had slightly higher error variance than respondents who were only given five tasks, while respondents with nine tasks had higher scale than respondents with five tasks. Brouwer et al. (2009) used mixed logit models in the context of a five task and three alternative experiment, finding increasing scale parameters consistent with learning effects, a suggestion that was confirmed by an analysis of respondent reported certainty measures.

Finally in a particularly important contribution, Bateman et al. (2008a), in a contingent valuation (CV) context, investigated the formation and nature of preferences by addressing an issue of particular importance to the valuation of low experience goods, namely the speed at which individuals can form stable preferences for relatively novel goods presented in unfamiliar markets. They developed a new approach, called learning design contingent valuation, to eliciting stated preferences for non-market goods. They found evidence of both institutional learning and value learning in repeated responses to CV questions. Valuations of an initial good exhibited typical anomalies, namely inconsistencies between single and double-bounded valuations of that good and anchoring effects. Analysis of trends in both within-good valuation differences and in anchoring showed significant reductions in both anomalies as repeated valuations are made. Indeed by the time respondents had undertaken a number of CV valuations both anomalies completely disappeared.

The evidence from the literature across a number of disciplines is clearly not conclusive, with the majority of papers showing either a lack of fatigue effects or even the presence of some learning effects, be it a context of unfamiliar or familiar markets. Most papers are also just based on a single dataset, often with a very limited number of choice tasks. However, the evidence from Bradley & Daly (1994) has been used so extensively that revisiting the issue seems appropriate. Indeed some recent work (cf. Brownstone et al., 2010) went as far as suggesting that model results are not reliable if they fail to account for the increasing error as a respondent progress through a stated choice experiment. While this paper wishes in no way to discredit the work of Bradley & Daly (1994), evidence from a single dataset is clearly not sufficient to justify making use of short surveys. It is also important to recognise the potential benefits of multiple choice tasks. As an example, Plott (1996) suggested that respondents may discover their true preferences through a learning process, and such learning process is expected to change preferences and thus parameter estimates in discrete choice experiments.

The topic of this paper is thus to revisit the issue of respondent fatigue in repeated choice settings. With a view to allowing us to reach more general conclusions than previous work, we look at a number of data sets collected in various countries and in different contexts, to investigate the extent of any systematic relationship between error variance (or scale) and choice set sequence. A key characteristic of the datasets is that the order of choice tasks is randomised across respondents, thus breaking the correlation between scale variation across tasks and attributes of the tasks – it is not clear that this has been the case in all (or even most) previous studies looking at fatigue effects. Additionally, while the majority of existing work has focussed solely on the investigation of scale differences across choice tasks, we offer a more comprehensive testing framework that also studies

differences in relative sensitivities across tasks, and also extend our analysis beyond simple Logit models to Mixed Logit structures.

Although there is evidence of some amount of difference in error variance, we were surprised to find that the differences were often small and had little influence on substantive model results. Crucially, there is no general decreasing trend in scale. Encouragingly from the perspective of the growing reliance on SC surveys with a large number of choice scenarios (up to 16 in our data sets), there is surprisingly little evidence of respondent fatigue. On the other hand, the study of the relative sensitivities across choice tasks suggests the possibility of learning of true preferences as a respondent proceeds through the survey. In conjunction, these findings suggest that analysts can capture choices for a large number of scenarios for each respondent (or at least larger than commonly assumed), giving respondents ample time to express their true preferences without any undue impacts of respondent fatigue.

The remainder of this paper is organised as follows. In the next section, we outline the empirical testing framework used in our analysis. This is followed in Section 3 by a discussion of the different datasets used in our empirical work. Section 4 presents model results, and finally, Section 5 offers a brief summary and conclusions.

2. Framework for empirical tests

This section briefly outlines the framework used for the empirical tests conducted for this study.

A total of six tests are carried out for each dataset, split into two groups of three tests. The first three tests relate to models estimated jointly on the data for all choice tasks, while the second set of three tests relate to models estimated separately for individual choice tasks. The first group of three tests thus relates to the work of Bradley & Daly (1994) and is concerned with studying the impact of allowing for choice task specific scale parameters. The second group of three tests makes of use of choice task specific models, thus also allowing for additional differences in relative sensitivities.

The first two tests make use of both multinomial logit (MNL) and mixed multinomial logit (MMNL) models, using a linear-in-attributes specification of the utility function, where, for the MMNL models, we made use of lognormal distributions for the random parameters. For the third test, we restricted ourselves to the MNL results, for reasons discussed below. The same applies to the three tests in the second group, as the small sample sizes would not have permitted the reliable estimation of choice task specific MMNL models. The MMNL models were specified with inter-respondent heterogeneity, thus recognising the repeated choice nature of the data. In both MNL and MMNL models, the panel specification of the sandwich error estimator was used to once again recognise the repeated choice nature of the data (cf. Guilkey & Murphy, 1993, and Daly & Hess, 2010).

We will now look at the tests in turn.

2.1. Scale differences across choice tasks

Three tests are conducted on each data set to investigate the effect of allowing for scale differences across choice tasks. The tests make use of information from two separate models. The first is a base model that does not allow for any differences across choice tasks. In particular, let U_{jnt} be the utility of alternative j for respondent n in choice task t (t = 1, .. T). In the base model, we then simply have that

$$U_{jnt} = V_{jnt} + \varepsilon_{jnt} = f(\beta, x_{jnt}) + \varepsilon_{jnt}, \quad (1)$$

i.e. the utility is composed of a *deterministic* component $V_{jnt,}$ and a random component ε_{jnt}, where this follows a type I extreme value distribution, with errors distributed identically and independently (*iid*) across tasks. The *deterministic* component V_{jnt} is given by $f(\beta,x_{jnt})$, where β and x_{jnt} are a vector of coefficients to be estimated and a vector of observed attributes respectively, and where the functional form for $f()$ depends on model specification. With the *iid* distribution of errors, and the constant β across tasks, any systematic differences in utilities across choice tasks are thus solely a function of differences in observed attributes.

In the second model, we allow for scale differences across choice tasks, by rewriting Equation (1) as:

$$U_{jnt} = \alpha_t f(\beta,x_{jnt}) + \varepsilon_{jnt}, \qquad (2)$$

where an appropriate normalisation is required, e.g. setting $\alpha_1=1$. With errors still specified as *iid* extreme value, an increase in α_t for a specific choice task equates to increased weight for the *deterministic* component of utility for choice task *t*, i.e. reduced variance for the error term. This specification is formally equivalent to the Nested Logit specification used by Bradley & Daly (1994) and others, but is based on the estimation of models with non-linear utility functions, where we use both MNL and MMNL models.

2.1.1. Test 1.1: Model fit impacts

Our first criterion is a likelihood ratio test[1] comparing the base model with the model making use of choice task specific scale parameters. In other words, with LL_{base} and LL_{scale} giving the log-likelihood obtained with models based on Equation (1) and Equation (2) respectively, we use $-2(LL_{\text{base}}-LL_{scale})\sim\chi^2_{T-1.}$ Here, the degrees of freedom for the test are equal to *T-1*, i.e. the number of estimated scale parameters in model 2. This test is carried out for both MNL and MMNL models.

2.1.2. Test 1.2: Evidence of scale differences

As a next step, we study the estimates for the choice task specific scale parameters α_t, looking for differences in scale across choice tasks, through comparison with the (normalised) scale for the first task, α_1. In particular, we look for any signs of trends across choice tasks in the values for the scale parameters, with reductions (i.e., higher unobserved variance) suggesting fatigue and increases (i.e., lower unobserved variance) suggesting learning effects. Once again, this test is carried out for both MNL and MMNL models.

2.1.3. Test 1.3: Impact on relative parameter estimates

As a final step, we investigate whether allowing for potential scale differences across choice tasks has any impact on substantive model results in the form of relative sensitivities, focussing on willingness-to-pay (WTP) indicators, and in particular also the significance levels of any differences between the models with and without choice task specific scale parameters. This test is only carried out for the MNL models so as to avoid the added difficulty of also studying differences in the implied heterogeneity in the WTP indicators, where only limited effort had gone into appropriate specifications of the distributions, using Lognormal distributions throughout.

2.2. Choice task specific models

Our second set of three tests move away from the assumption that any impacts of survey duration are restricted to scale differences and thus looks at choice task specific models that allow for

[1] For the MNL models, this test is approximate as it does not recognise the repeated choice nature of the data.

variation also in relative sensitivities. Given sample size limitations, these tests are restricted to MNL models only.

In particular, we estimate a separate model for each choice task. It can be seen that the combined results for these T models are equivalent to those that would be obtained from a single model in which we rewrite the specification in Equation (1) as:

$$U_{jnt} = f(\beta_t, x_{jnt}) + \varepsilon_{jnt}, \quad (3)$$

i.e. using choice task specific parameters.

The results from these choice task specific models could again highlight the presence of specific trends, such as respondents focussing more or less on specific coefficients as the survey progresses. As an example, one school of thought would be that respondents initially focus on their most important attributes while they get used to the experiment, while another view would be that respondents in time focus on just such a subset. Both phenomena could be seen as evidence of learning effects, while the second could similarly be interpreted as fatigue.

2.2.1. Test 2.1: Model fit impacts

Our first test in this group looks at the impact on model fit of allowing for choice task specific models. In particular, we compute a combined model fit (made up of the fits of the individual models), where this is given by $LL_{combined}=\sum_t LL_t$, where LL_t gives the log-likelihood for the model estimated on the data for choice task t only. We then conduct likelihood-ratio tests for this combined model against the model assuming complete homogeneity across task, i.e. $-2(LL_{base}-LL_{combined})\sim\chi^2_{(T-1)p}$, and the model allowing for scale differences but maintaining an assumption of homogeneity in relative sensitivities, i.e. $-2(LL_{scale}-LL_{combined})\sim\chi^2_{(T-1)p-T+1}$.

2.2.2. Test 2.2: Trends in model fit

Next, we look at any trends in model fit across the choice task specific models, i.e. a study of LL_t for $t=1,...T$, noting that this is strongly related to the scale difference test in section 2.1.2., with higher scale equating to greater weight for the *deterministic* component of utility, and hence more extreme choice probabilities (with the choice probability for the alternative with the highest *deterministic* utility increasing as scale increases). Here, we also look for correlation between the choice task specific log-likelihood contributions and the task number.

2.2.3. Test 2.3: Evidence of trends in relative importance of different attributes

The model making use of choice task specific scale parameters allows for scale differences across choice tasks. The choice task specific models allow for such scale heterogeneity but also additional heterogeneity in the relative importance of difference attributes across tasks. We first compare the degree of heterogeneity in the model with scale heterogeneity only to the degree of heterogeneity in the choice task specific models, on the basis of the coefficient of variation. In the model with choice task specific scale parameters, this will, for coefficient k (i.e. β_k), be given by $cv_{k,scaled}=\sqrt{(var(\alpha_1\beta_k,..., \alpha_T\beta_k))}/\beta_k$, i..e the heterogeneity will clearly be constant across coefficients. On the other hand, in the model based on Equation (3), we would have $cv_{k,combined}=\sqrt{(var(\beta_{k1},..., \beta_{kT}))}/(\mu(\beta_{k1},..., \beta_{kT}))$, where β_{kt} gives the value for coefficient k in choice task t. If any differences across choice tasks could be explained by the choice task specific scale parameters, we would have that $cv_{k,combined}/cv_{k,scaled}=1$. If on the other hand this ratio is greater than one, we have evidence of further heterogeneity, net of scale differences. Where there is such additional heterogeneity, it is of interest to look for any trends. For this purpose, it is necessary to first disentangle the two types of heterogeneity, where this can be achieved by factoring out the scale differences in the choice task specific models, through dividing each coefficient in each of the T models by the appropriate scale

parameter from the model with choice task specific scale parameters. In other words, we would have that $\beta_{kt,net}=\beta_{kt}/\alpha_t$. The resulting coefficients are *net* of scale differences, and we can look for correlation between coefficients and tasks, where we also compute a *t*-ratio for such correlation. If we obtain significant positive correlation between task numbers and $<\beta_{k1},..., \beta_{kT}>$, then this equates to decreasing sensitivity as the survey progresses if β_k is a *negative* coefficient, and increasing sensitivity as the survey progresses if β_k is a *positive* coefficient. The opposite reasoning applies to negative correlation.

3. Data

In this section, we present the different datasets used in our analysis. These datasets were collected in different countries (UK, Denmark, Australia, USA) and made use of different choice scenarios in terms of number of alternatives (two or three) as well as number of attributes (from two to six), while some additional differences arise in the number of choice tasks faced by each respondent (ranging from eight to sixteen). Finally, three of the surveys were conducted as computer aided personal interviews (CAPI), while a fifth one (fungibility study) was conducted as an online survey. The Atlanta survey used a mixture of these two data collection methods. These differences across surveys allow us to reach more general conclusions.

The crucial common factor across all surveys is that for each respondent, the order of the *T* tasks was randomised. This ensures that any scale differences retrieved in the analysis should be free of effects of the specific make up of individual choice tasks; in other words, there should be no correlation attribute levels (and potentially resulting task complexity) and the scale effects retrieved in our analysis. Indeed, if the same ordering had been used across respondents, and if say the first task had always been *easier*, then some deterministic patterns in scale heterogeneity could have been expected. As already alluded to earlier, it is not clear that this requirement was met in all or even most previous studies looking at respondent fatigue.

3.1. Atlanta toll road study

The first case study makes use of data collected for a toll road study in Atlanta (see Hess et al., 2008 for more details). In each choice task, a respondent was faced with three alternatives; driving in the existing untolled lanes (general lanes), driving in a tolled lane (managed lanes), or carpooling in the managed lanes in return for a reduced toll. The three alternatives were described by two attributes, namely travel time and toll (zero for general lanes). For each respondent, data was collected from eight choice tasks.

Three different samples were collected for this study, with differences in the underlying design approach. In the first sample of 1,563 respondents, an orthogonal design was used, with the eight tasks for each respondent being drawn at random from the overall design, i.e., making use of random blocking. In the second sample of 1,146 respondents, the same underlying design was used, but the eight tasks for each respondent were obtained with orthogonal blocks. For the final sample of 1,110 respondents, a D-efficient design was used, once again with orthogonal blocking. There were some socio-demographic differences across the three samples, as discussed by Hess et al. (2008), and these are partly reflected in the differences in the WTP measures across the three samples. In each of the three samples, the data was collected via an internet based survey.

3.2. Fungibility data

The second case study used data collected by Orr et al. (2010) in a study looking at the fungibility of monetary valuations in a transport safety context. Specifically, the survey looked at the relative sensitivities to rail travel time, costs, and safety (number of accidents). Each respondent faced three different binary SC experiments, trading time against cost, time against safety, and safety against cost. Each experiment made use of five choice tasks, equating to a total of 15 choices per

respondent, where a D-efficient design was used. The order of the three experiments was randomised across respondents, with the same number of respondents in each of six orderings, and the order of choice tasks was also randomised within each experiment. Finally, each respondent was additionally presented with three corresponding contingent (CV) valuation exercises involving the binary comparisons, where half the respondents were given the CV experiments first, with the other half being given the SC experiments first. A final sample of 397 respondents was obtained for this study; the slight loss of balance across the twelve different subgroups was so small as to be inconsequential.

3.3. Danish Value of Time data

The third case study makes use of data from a binary unlabelled abstract choice experiment conducted in Denmark, with the two alternatives being described by travel time and travel cost. The attribute combinations were based on a manual design encouraging trading between the two attributes. For further details on this survey, see for example Fosgerau (2006). Each respondent was faced with up to eight choice tasks, and for the present study, we made use of a sample of 3,633 observations from 472 commuters, and a sample of 13,387 observations from 1,725 non-commuters.

3.4. Sydney M4 data

Our fourth case study makes use of data collected in a three alternative route choice experiment in Sydney (cf. Hensher & Rose, 2005), making use of a D-efficient design. Of the three alternatives, the first corresponded to a reference trip where the attributes for that alternative were kept invariant across the sixteen choice tasks faced by the respondent. The three alternatives were described in terms of five attributes, namely free flow time, slowed down time, running costs, tolls, and travel time variability. For the present study, we made use of samples of 3,792 observations from 237 commuters, and 3,280 observations from 205 non-commuters.

3.5. Second Australian dataset

Our fifth case study makes use of a survey very similar to that from the fourth case study, but collected in a different Australian city in 2005, and making use of an additional travel time component, described as crawl time. For further details on this survey, see the recent application in Hess et al. (2010). A D-efficient design was used once again, and each respondent was faced with sixteen choice tasks. For the present analysis, we made use of a sample of 4,864 observations from 304 commuters.

4. Empirical analysis

We now proceed with the discussion of the empirical results, taking each dataset in turn. Separate tables are used for the MNL results for each dataset (or subsample), giving us Table 1 to Table 7. The results for the two MMNL tests are summarised in a joint table across all datasets (Table 8).

4.1. Atlanta toll road study

The model specification for the Atlanta data made use of alternative specific constants (ASC) for the first two alternatives (GM & ML), along with marginal utility coefficients for time and tolls, and an additional dummy (penalty) term if carpooling (i.e. alternative 3) meant increasing vehicle occupancy by two people (OCC). For the MMNL models, the two marginal utility coefficients were allowed to vary across respondents. We will now look at the results from the various tests in turn, across the different subsamples.

The results for the first three tests using MNL models on the Atlanta data are summarised in Table 1. In test 1.1., we note that the use of choice task specific scale parameters leads to improvements in log-likelihood by 1.32 units for the first sample, 7.12 units for the second sample, and 12.09 units for the third sample. Each step comes at the cost of seven additional parameters; we see that this improvement is not significant at any reasonable level of confidence for the first sample, while it is significant at the 95% level of confidence for the second sample and the 99% level for the third sample. On the other hand, the MMNL results in Table 8 show significant improvements in fit only for the first sample.

Given the results of the likelihood ratio tests, we would thus expect an absence of scale heterogeneity for the MNL results for first sample, a finding that is confirmed in test 1.2., with no scale parameter being significantly different from 1 (the first choice task being used as the base). The evidence of a small but significant improvement in MNL model fit for the second sample is consistent with the findings from test 1.2., which show some significant differences in scale across choice tasks. Crucially however, the scale parameters for all choice tasks are greater than that for the first task, and the fluctuation we observe after this first task shows no sign of any clear trend indicating either learning or fatigue effects. Similarly, the MNL results for the third sample do indeed suggest some significant scale differences, but the trend is one of increases, especially after the fourth choice task, suggesting evidence of learning rather than fatigue. For the MMNL results, only a handful of scale parameters (across the three models) obtain modest levels of statistical significance for t-ratios against a base value of 1, but crucially, no evidence of significant reductions in scale are observed. The differences in results between MNL and MMNL are somewhat to be expected, given that the latter now explains part of the error through the heterogeneity specification. As alluded to earlier, there are small differences in socio-demographics between the three samples, primarily relating to a greater share of respondents travelling for work reasons in the second and third group (and hence travelling more regularly), where these groups also show somewhat higher income. In part, these variations could explain the differences in results, and especially the greater propensity for learning in these two groups in the MNL model. See also discussions by Ladenburg and Olsen (2008) in this context.

Turning our attention next to the impact that allowing for choice task specific scale parameters has on substantive MNL model results (i.e. test 1.3), we observe essentially no differences in any of the four ratios between the base model and the model with choice task specific scale parameters, across the three samples.

The results for the second set of three tests on the Atlanta data are summarised in Table 2. Looking first at test 2.1, we observe that for the first sample, the improvement in fit obtained from using choice task specific models is not statistically significant when compared to the base model and the model with choice task specific scale parameters. On the other hand, in the second and third sample, the use of choice task specific models leads to improvements over the base model and the model with choice task specific scale parameters that are significant beyond the 99% level of confidence.

The results from test 2.2 for the first sample are in line with the lack of improvement obtained with the choice task specific models and the absence of significant differences in scale across choice tasks. Indeed, there is no evidence of any significant variation in model fit across the eight choice task specific models. On the other hand, we obtain some evidence of a gradual improvement in the case of the second sample, with a much stronger effect in the case of the third sample. These findings are line with the increases in scale observed in test 1.2 for these two samples.

Looking finally at test 2.3, we note that while for the first sample, there are increases in heterogeneity compared to the model with choice task specific scale parameters; this is not surprising given the almost complete lack of scale differences in this sample. However, the

fluctuation seems to be almost completely random, with two exceptions: we see a decrease in the estimates for the general lanes constant; and some evidence of increased toll sensitivity as the experiment progresses. In the second sample, we see increased heterogeneity for all parameters except the managed lanes constant, but this heterogeneity net of scale shows no trends across the eight choice tasks. In the third sample, we once again see evidence of heterogeneity on top of the scale variations, with a suggestion of decreasing sensitivity to increased occupancy (remembering that the estimate is negative), and increasing sensitivity to tolls and travel time as the experiment progresses.

4.2. Fungibility study

The specification for the models estimated on the fungibility data made use of three linear coefficients, for travel time (minutes), cost (£), and the number of accidents (in 1000s), where these were allowed to vary across respondents in the MMNL models. To allow us to separate out the effects of survey length, the models allowed for scale differences between the three different SC experiments, and also differences in scale depending on whether respondents took part in the CV experiment before or after the SC experiment. The results showed no significant scale differences depending on this latter ordering, while we observed that the scale for the time *vs.* money experiment is higher than for the cost *vs.* safety experiment, which in turn has higher scale than the time *vs.* safety experiment.

The MNL results for this study are summarised in Table 3. We obtain a base model fit of -3,429.50, which increases to -3,418.87 when using choice task specific scale parameters; this improvement by nine units for 14 parameters is significant only at the 90% level (test 1.1). On the other hand, the improvement in the MMNL case is significant at high levels of significance. There is no evidence of fatigue when studying the evolution of scale parameters in the MNL model, and in fact, there is some evidence of increases in scale which would suggest possible learning effects (test 1.2). In the MMNL models, we note an initial drop in scale, but this is then followed by rather random variation. We also note no impact on the MNL WTP estimates as a result of allowing for choice task specific scale parameters (test 1.3). We observe improvements in the per-observation contribution to the log-likelihood function as the experiment progresses, in line with the results for the scale differences (test 2.2).

The results from test 2.1 show that using choice task specific models leads to significant gains in model fit over the unscaled and scaled base models, suggesting the presence of variations in sensitivities across choice tasks. Finally, the results for test 2.3 show evidence of attribute specific heterogeneity (i.e. on top of the scale differences) for each of the six parameters, where this is significant for all parameters except for the scale parameter associated with giving the CV experiment first. There is evidence of increasing sensitivity to time, safety and cost, but by different degrees. This finding is interesting, and could be linked to *learning* of the true valuations, as discussed by Plott (1996). There is also some evidence of decreasing scale parameters for the cost *vs.* safety and the time *vs.* safety experiments – this would suggest that the differences in scale between the three types of experiment are especially pronounced later in the survey. Here, one could argue that this is evidence of reduced engagement with these more *difficult* trade-offs later in the survey, but this is not in line with the overall findings in terms of a lack of scale reductions.

4.3. Danish VOT data

The models estimated on the Danish VOT data made use of linear travel time (minutes) and travel cost (Øre, with 1 Danish Crown (DKK)=100 Øre) coefficients, along with a constant for first alternative, included after evidence of significant effects of reading left to right. The two marginal utility coefficients were allowed to vary randomly across respondents in the MMNL models. The

MNL results for this dataset are summarised in Table 4, with separate models for commuters and non-commuters.

The base model estimated on the commuter data obtained a log-likelihood of -2,404.02, which improves to -2,396.33 when using choice task specific scale parameters, where this increase in eight units for seven additional parameters is significant at the 95% level. However, from the results for test 1.2, it becomes clear that while there is some variation in estimated scales, there are no clear trends, and none of the differences are statistically significant.

For the models estimated on the non-commuter data, the incorporation of choice-task-specific scale parameters leads to a highly significant improvement in model fit from -8,898.48 to -8,877.16, at the cost of seven additional parameters. This is consistent with the findings for test 1.2 which show a very sharp drop in scale when moving from the first to the second choice task, followed by an increase and greater stability from there onwards. A closer inspection of the results (results available on request) allows us to link this finding directly to the values for the ASC for the first alternative. The estimate for this constant is so high and dominant in the first task that the travel time coefficient is not in fact statistically significant – this also explains the much lower valuation of travel time (VTT) for the first task. After the first task, the constant rapidly drops in value. This is again an indication of strong reading left to right effects for the first choice task.

For the MMNL models (see Table 8), we observe highly significant gains in fit for both samples when incorporating scale differences across choice tasks. Alongside the drop in scale already mentioned for the MNL model for non-commuters, we see heightened scale later on in the experiment for both samples.

Test 1.3 shows that the inclusion of choice task specific scale parameters has no impact on the estimated valuation of travel time (DKK/hour), in either of the two segments. From the results for test 2.1, we can see that the estimation of choice task specific models leads to highly significant gains in model fit when compared to both the unscaled and the scaled base model. Test 2.2 shows some fluctuation in model fit across the eight choice task specific models. However, there is no statistically significant trend. The main observation relates to the higher fit for the first choice task, which is in line with the above discussions on the importance of the constant in the first choice task, with a much higher rate of choosing the left alternative (leading to choices that are *easier* to model). This heterogeneity in the constant is also reflected in the results for test 2.3. We see evidence of significant attribute level heterogeneity for the constant, net of the scale difference, where, importantly, there is clear evidence of decreasing values for the constant as the experiment progresses. There is also evidence of increasing time sensitivity for non-commuters as the experiment progresses, while the other changes are of lesser importance.

4.4. Sydney M4 data

The models estimated on the Sydney data made use of ASCs for the first two alternatives, along with linear coefficients for the free flow time, slowed down time, and travel time variability coefficients (all in minutes), and the running costs and toll coefficients (in AUD). The MNL results are summarised in Table 5 for the commuter sample and Table 6 for the non-commuter sample.

Looking first at test 1.1 for the commuter sample MNL models, we see that the base model obtains a log-likelihood of -2,854.28 which rises to -2,846.23 when incorporating the fifteen choice task specific scale parameters, an increase that is only significant at the 62% level. For the non-commuter models however, the incorporation of choice task specific scale parameters leads to an increase in the MNL log-likelihood from -2395.88 to -2378.15 at the cost of fifteen additional parameters, an improvement that is significant at the 99% level.

The results for test 1.2 for the commuter sample show no evidence of fatigue for the MNL model, and in fact possibly suggest some evidence of learning effects, although none of the differences are statistically significant. For non-commuters, we observe some variations in scale across choice tasks in the MNL models, with an initial drop followed by a renewed increase, and a definite absence of a consistent downwards trend. There is no evidence in either sample of any impact on the relative sensitivities as a result of allowing for choice task specific scale parameters (test 1.3).

In the MMNL models, we see significant improvements in fit in both samples as a result of incorporating choice task specific scale parameters. Crucially, there is no evidence of fatigue.

The estimation of choice task specific models leads to significant gains in model fit for both samples when compared to the unscaled and scaled base models (test 2.1). Interestingly however, the results in terms of choice task specific log-likelihood (test 2.2) are somewhat contrary to the findings in terms of significant scale variations, with evidence of a significant increasing trend for the commuter sample, while the variation for the non-commuter sample seems to be more random.

The results for test 2.3 show significant variation in the constants net of the scale heterogeneity, where, with the positive estimates for the two constants, there is evidence of decreasing values for the constant, albeit that this is not statistically significant. For the commuter sample, we see additional variation especially for the slowed down time and toll coefficients, where the only statistically significant trend is however an increasing cost sensitivity as the experiment progresses. For non-commuters, we also see significant additional heterogeneity especially for toll, but no indication of any clear trends in the evaluation of the parameter estimates.

4.5. Second Australian dataset

The model specification for the second Australian dataset is very similar to that used in the Sydney case study, with the addition of a coefficient for crawl time. The MNL estimation results in Table 7 show that that the inclusion of choice-task-specific scale parameters leads to an improvement in log-likelihood from -2,668.39 to -2,644.62, which, at the cost of fifteen additional parameters, is significant at the 99% level (test 1.1). Once again, the actual trend is not completely clear, but there is possibly some evidence of learning, and clearly no evidence of fatigue (test 1.2), with once again consistent results for the changes in model fit (test 2.1), which show some evidence of significant increases in choice task specific log-likelihood as the experiment progresses. Again, there is no evidence of any impact on the relative sensitivities as a result of allowing for choice task specific scale parameters. In the MMNL models, we again observe a significant increase in model fit when incorporating scale differences, where these provide no evidence of respondent fatigue.

The results for test 2.1 also show additional gains by making use of choice task specific models, suggesting the presence of differences across tasks that cannot be explained solely through the use of choice task specific scale parameters. Here, the results from test 2.3 show high levels of additional heterogeneity especially for the two constants and for the travel time variability coefficient. However, there are no clear trends for any of the estimates, except maybe the cost coefficient, which suggests reducing cost sensitivity as the experiment progresses.

5. Conclusions

This paper has attempted to provide a more reliable answer to the persisting question as to the presence of fatigue effects in SC surveys. Specifically, we have presented evidence from five different SC surveys, some of them split into several subsamples. We have made use of a comprehensive empirical framework employing both logit and mixed logit models. The degree of scale heterogeneity across choice tasks varies substantially across the different datasets, with in some cases no evidence of any differences over the presented sequence of choices. There seems to be more evidence of scale differences in the MMNL models. More crucially however, in those

datasets where significant differences are observed, there is no clear evidence of any consistent decrease in scale over the duration of the experiment, with the opposite applying in some cases, indicating the possible existence of learning effects. A summary plot showing the scale variations in shown in **Error! Reference source not found.**, which shows that overall, the variations are more amplified in the MNL models, but that there is no conclusive evidence relating to respondent fatigue, and more suggestions relating to learning effects. As already mentioned, the visible drop observed for the Danish non-commuter sample is a result that we attribute to the high value for the alternative specific constant in the first task. It should be noted that there is also evidence of higher weight for the constants in early tasks in other datasets, as reflected for example in the decreasing importance of the constants in the M4 commuter data as the experiment progresses.

Additionally, in these data sets, while parameter estimates show variation on top of scale differences, this variation is mostly random, although there are trends in some cases which could be possible evidence of learning effects, as previously discussed by Plott (1996). Here, Figure 2 shows trends in the MNL WTP measures, with the WTP in all datasets normalised to 1 for the first task. Some of the datasets show very stable results, while others, most notably the trade-off between slowed down time and running costs in the second Australian dataset, change quite substantially. On average, there are more increases than decreases, possibly suggesting initial cost aversion followed by learning of the trade-offs. This point also relates to the fact that overall, there is more evidence of learning (in terms of increasing scale) than fatigue. The larger constant for the first task can be linked to this in that it could signify that respondents simplify the first task by choosing the first alternative. This is also supported in discussions by Carlsson et al. (2011) who suggest that the first choice is the most difficult for respondents.

The findings in this paper are consistent with several other studies over the years, as outlined in the introduction. The difference is that our work is based on multiple datasets and makes use of a comprehensive testing framework, allowing us to more easily generalise the results. On balance, the weight still given to the early evidence in Bradley & Daly (1994) seems unjustified. The question still arises though as to why the findings in that paper were so substantially different. One possible reason lies in the nature of the survey. Rather than being based on a design that gave each respondent a fixed set of choice tasks, the underlying design produced a set of nine alternatives for each respondent, drawn randomly from the full set of possible attribute combinations. The respondent was then presented with randomly produced binary choices from these nine alternatives up to a point where a full preference ordering could be inferred. The possible repeated occurrence of an alternative is one possible reason for the evidence of fatigue (or boredom) in that data. In this context, an interesting avenue for future work would be to attempt to link the different observations on fatigue to characteristics of the sample as well as the survey. The present work made use of data from several substantially different surveys, with no obvious link between the findings and these differences in terms of sample as well as survey design. Similarly, there is no obvious link between the data collection method and the results. This however is an area that deserves further investigation. One argument would be that fatigue might set in earlier in self administered surveys. On the other hand, self administered surveys can be completed at a time that is convenient to the respondent (unlike interviewer led surveys) and this may again have a positive impact. Finally, a case could be made for reduced risk of fatigue when respondents are familiar with the choice scenarios. This could in part explain the greater scope for learning in the second and third Atlanta sample which involve more regular travellers. The question also arises whether the somewhat random fluctuations for the fungibility study are related to inexperience with the specific trade-offs used in that survey. It must also be recognised that differences exist between data sets that cannot be fully explained.

The evidence in the present paper, and that from a number of other applications described in the introduction, should serve to somewhat ease the concerns about respondent fatigue, and the above

discussion possibly explains the discrepancies with the work by Bradley & Daly (1994). While testing for fatigue and learning effects is still good practice, it should also be noted that it is not practical to work with scale differences in a final model used for implementation. Indeed, the marginal effect of a given attribute in choice set *t* would then be given by $\mu_t\beta$, meaning that the *t*-ratios will vary across tasks, leading to complications in producing a single measure of robustness for a given effect. Finally, there is clear evidence across all our case studies that even in the presence of significant scale differences across choice tasks, the impact of these differences on the WTP measures is minimal at the most.

Acknowledgement

This work described in this paper was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the project "Estimation of travel demand models from panel data", grant EP/G033609/1. The first author also acknowledges the support of the Leverhulme Trust in the form of a Leverhulme Early Career Fellowship. The authors are grateful to Phani Kumar Chintakayala for background research; and to two anonymous referees for comments on an earlier version of this paper.

References

Adamowicz, W., Boxall, P., Williams, W., and Louviere, J.J. (1998), Stated preference approaches for measuring Passive use values: choice experiments and contingent valuation *Amer. J. Agr. Econ. 80 (February 1998): 64-75.*

Bateman, I., Carson, R., Dupont, D., Day, B., Louviere, J.J., Morimoto, S., Scarpa, R., and Wang, P. (2008), Choice set awareness and ordering effects in choice experiments, 16th Annual Conference of the European Association of Environmental and Resource Economics (EAERE), Gothenburg, Sweden; 25-28 June, 44pgs.

Bateman, I., Burgess, D., Hutchinson, W.G., and Matthews, D. I. (2008a) Learning design contingent valuation (LDCV): NOAA guidelines, preference learning and coherent arbitrariness, *Journal of Environmental Economics and Management,* 55(2):127-141.

Bech, M., Kjaer, T., and Lauridsen, J. (2010), Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment, *Health Economics*, February 2010.

Bliemer, M.C. and Rose, J.M. (2009) Designing Stated Choice Experiments: The state of the Art, in Kitamura, R., Yoshi, T. and Yamamoto, T., *The Expanding Sphere of Travel Behaviour Research, Selected Papers from the 11th International Conference on Travel Behaviour Research*, Ch25, 495-498.

Bradley, M and Daly, A. (1994) Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data, *Transportation*, 21, 2, 167-184

Brazell, J.D. & Louviere, J.J. (1996), Helping, Learning, and Fatigue: An Empirical Investigation of Length Effects in Conjoint Choice Studies, Department of Marketing, The University of Sydney.

Brouwer, R., Dekker, T., Windle, J.R.J. (2009), Choice certainty and consistency in repeated choice experiments. *Environmental and Resource Economics*. 46 (1) 93-109.

Brownstone, D., Hansen, M. & Madanat, S. (2010), Review of "Bay Area/California High-Speed Rail Ridership and Revenue Forecasting Study", UC Berkeley Research Report UCB-ITS-RR-2010-1, Institute of Transportation Studies, University of California, Berkeley, CA.

Carlsson, F., Mørkbak, M.R. & Olsen, S.B. (2011), The first time is the hardest: A test of ordering effects in choice experiments, paper presented at the Second International Choice Modelling Conference, Oulton Hall, Leeds.

Caussade, S., Ortúzar, J. de D., Rizzi, L., and Hensher, D.A. (2005), Assessing the Influence of Design Dimensions on Stated Choice Experiment Estimates. *Transportation Research B, 39: 621–640.*

Daly, A. & Hess, S. 2010, Simple methods for panel data analysis, paper presented at the European Transport Conference, Glasgow

Fosgerau, M. (2006), Investigating the distribution of the value of travel time savings. *Transportation Research Part B* 40 (8), 688-707.

Guilkey, D.K. and Murphy, J.L. (1993), "Estimation and testing in random effects probit model", Journal of Econometrics 59, 301-317.

Hanley, N., Wright, R.E., and Koop, G. (2002) Modelling Recreation Demand Using Choice Experiments: Climbing in Scotland, *Environmental and Resource Economics* 22: 449–466.

Hensher, D.A. and Rose, J. (2005) Respondent Behavior in Discrete Choice Modeling with a Focus on the Valuation of Travel Time Savings, *Journal of Transportation and Statistics*, 8 (2), pp. 17-30.

Hensher, D. A., 2010. Attribute processing, heuristics, and preference construction in choice analysis. In Hess, S. & Daly, A (eds), *Choice Modelling: the state-of-the-art and the state-of-practice;* Proceedings from the Inaugural International Choice Modelling Conference. Emerald, ch. 3, pp.35-70.

Hess, S., Smith, C., Falzarano, S. & Stubits, J. (2008), Measuring the effects of different experimental designs and survey administration methods using an Atlanta Managed Lanes Stated Preference survey, *Transportation Research* Record, 2049, pp. 144-152.

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.

Holmes, T. and Boyle, K.J. (2005), Dynamic Learning and Context-Dependence in Sequential, Attribute-Based Stated-Preference Valuation Questions, *Land Economics* 81: 114-126.

Hu, W. (2006) Effects of Endogenous Task Complexity and the Endowed Bundle on Stated Choice, Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Long Beach, California, July 23-26, 2006

Ladenburg J., and Olsen SB (2008) Gender-specific starting point bias in choice experiments: evidence from an empirical study. J Environ Econ Manag 56(3): 275–285

McNair, B., Bennett, J., and Hensher, D. A. (2010), Strategic response to a sequence of discrete choice questions, 2010 Conference of the Australian Agricultural and Resource Economics Society, Adelaide

Orr, S., Hess, S. & Sheldon, R. (2010), Fungibility of monetary valuations in a transport context: an empirical investigation of the transferability of willingness to pay measures, paper presented at the European Transport Conference, Glasgow

Phillips, K.A., Johnson, F.R., and Maddala, T. (2002), Measuring What People Value: A Comparison of 'Attitude' and 'Preference' Surveys, *Health Services Research,* 37: 1659-1679.

Plott, C.R. (1996), Rational Individual Behavior in Markets and Social Choice Processes: The Discovered Preference Hypothesis, in *Rational Foundations of Economic Behavior*, K. Arrow, E. Colombatto, M. Perleman, and C. Schmidt, eds., Macmillan, London.

Raffaelli. R, Notaro.S, Scarpa.R, Pihlens. D, Louviere. J. (2009) Valuing the external effects of Alpine transhumance: an application of the best-worst approach to rank ordered data, paper presented at the International Choice Modelling conference, Harrogate.

Risa Hole, A. (2004), Forecasting the demand for an employee Park and Ride service using commuters' stated choices, Transport Policy, vol. 11(4), pages 355-362.

Savage, S. and Waldman, D. (2008) Learning and Fatigue During Choice Experiments: A Comparison of Online and Mail Survey Modes, *Journal of Applied Econometrics*, 23, 351-371

Swait, J. and W.L. Adamowicz (2001), The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Switching Strategy, *Journal of Consumer Research,* 28: 135-148.

Table 1: Atlanta data: results for tests 1.1.-1.3.

		Sample 1				Sample 2				Sample 3			
TEST 1.1	LL base	-10,088.24				-5,985.65				-5,715.62			
	par	5				5				5			
	LL scale	-10,086.92				-5,978.53				-5,703.53			
	choice tasks	8				8				8			
	LR-test	2.63				14.24				24.18			
	df	7.00				7.00				7.00			
	p-value	0.92				0.05				0.00			
TEST 1.2 (scale parameters)	Choice task	est.		t-ratio (vs. 1)		est.		t-ratio (vs. 1)		est.		t-ratio (vs. 1)	
	1	1		-		1		-		1		-	
	2	1.04		0.77		1.11		1.82		1		0	
	3	1.02		0.38		1.14		2.23		1.05		0.81	
	4	0.99		-0.21		1.07		1.14		1.05		0.83	
	5	1.02		0.36		1.2		2.97		1.14		2.13	
	6	0.99		-0.15		1.07		1.15		1.17		2.46	
	7	1.07		1.22		1.19		2.78		1.21		2.94	
	8	1		0		1.17		2.7		1.23		3.22	
TEST 1.3 (WTP)		time ($/hr)	general lanes vs car pool ($)	managed lanes vs car pool ($)	avoid occupancy increase by 2 ($)	time ($/hr)	general lanes vs car pool ($)	managed lanes vs car pool ($)	avoid occupancy increase by 2 ($)	time ($/hr)	general lanes vs car pool ($)	managed lanes vs car pool ($)	avoid occupancy increase by 2 ($)
	Unscaled est	11.45	9.03	5.14	0.57	8.84	7.39	4.22	0.54	8.12	6.66	4.17	1.11
	Unscaled t-rat	11.30	10.40	11.30	1.90	12.50	11.40	12.90	2.60	11.10	10.80	12.70	4.30
	Scaled est	11.44	9.02	5.13	0.56	8.81	7.30	4.18	0.54	8.09	6.58	4.14	1.09
	Scaled t-rat	11.30	10.40	11.30	1.80	12.60	11.50	12.90	2.60	11.20	10.80	12.70	4.20
	t-rat (diff)	-0.01	-0.01	-0.01	-0.02	-0.01	-0.10	-0.08	0.00	-0.02	-0.09	-0.08	-0.06

Table 2: Atlanta data: results for tests 2.1.-2.3.

		Sample 1					Sample 2					Sample 3				
Test 2.1	LL combined	-10,070.35					-5,946.36					-5,673.81				
	par	40					40					40				
	LR-test (comb vs unsc)	35.79					78.59					83.62				
	df	35					35					35				
	p-value	0.43					0.00					0.00				
	LR-test (comb vs sc)	33.15					64.35					59.43				
	df	28					28					28				
	p-value	0.23					0.00					0.00				
Test 2.2 (trends in LL)	Task	per obs LL					per obs LL					per obs LL				
	1	-0.8057					-0.6903					-0.6710				
	2	-0.7968					-0.6616					-0.6782				
	3	-0.8007					-0.6349					-0.6709				
	4	-0.8129					-0.6728					-0.6623				
	5	-0.8105					-0.6115					-0.6299				
	6	-0.8134					-0.6674					-0.6188				
	7	-0.7897					-0.6161					-0.5865				
	8	-0.8132					-0.6342					-0.5941				
	correl (t-rat)	-0.17 (-0.43)					0.6 (1.83)					0.94 (7.03)				
	scaled model	-0.8067					-0.6521					-0.6423				
Test 2.3 (relative heterogeneity)		GL	ML	OCC	toll	time	GL	ML	OCC	toll	time	GL	ML	OCC	toll	time
	cv (scaled)	0.03	0.03	0.03	0.03	0.03	0.06	0.06	0.06	0.06	0.06	0.08	0.08	0.08	0.08	0.08
	cv (individual)	0.07	0.07	1.25	0.13	0.08	0.05	0.13	1.04	0.21	0.14	0.10	0.11	0.43	0.20	0.21
	rate of change	+165%	+159%	+4,801%	+424%	+204%	-18%	+129%	+1,695%	+255%	+134%	+22%	+44%	+450%	+161%	+162%
	corr between task & est. net of scale	-0.56	-0.37	0.14	-0.63	0.44	-0.51	-0.03	0.12	-0.49	-0.40	-0.06	0.25	0.68	-0.65	-0.67
	t-rat	-1.65	-0.97	0.34	-2.00	1.19	-1.44	-0.08	0.30	-1.38	-1.07	-0.14	0.64	2.26	-2.12	-2.24

Table 3: Results for fungibility data

TEST 1.1

LL base	-3,429.50
par	6
LL scale	-3,418.87
choice tasks	15
LR-test	21.27
df	14
p-value	0.09

TEST 1.2 (scale parameters)

Choice task	est.	t-ratio (vs. 1)
1	1.00	-
2	1.19	0.72
3	1.61	1.69
4	1.38	1.24
5	1.36	1.15
6	1.38	1.06
7	1.26	0.86
8	1.40	1.24
9	1.42	1.17
10	1.50	1.45
11	1.54	1.35
12	2.06	2.20
13	1.64	1.45
14	1.42	1.16
15	1.61	1.60

TEST 1.3 (WTP)

	time vs cost ($/hr)	safety vs cost ($/1000acc)
Unscaled est	6.12	3.8
Unscaled t-rat	13.1	8.9
Scaled est	6.23	3.89
Scaled t-rat	13.4	8.8
t-rat (diff)	0.17	0.15

TEST 2.1

LL combined	-3,369.19
par	48
LR-test (comb vs unsc)	120.62
df	42
p-value	0.00
LR-test (comb vs sc)	99.35
df	28
p-value	0.00

TEST 2.2

Task	per obs LL
1	-0.6086
2	-0.5884
3	-0.5476
4	-0.5764
5	-0.5792
6	-0.5563
7	-0.5834
8	-0.5799
9	-0.5708
10	-0.5608
11	-0.5600
12	-0.5002
13	-0.5494
14	-0.5672
15	-0.5584
correl (t-rat)	0.55 (2.37)
scaled model	-0.5741

Test 2.3 (relative heterogeneity)

	cv (scaled)	cv (individual)	rate of change	corr between task & est. net of scale	t-rat
cost	0.16	0.28	+76%	-0.75	-4.12
safety	0.16	0.49	+209%	-0.91	-7.98
time	0.16	0.53	+237%	-0.83	-5.47
Scale (CV first)	0.16	0.55	+248%	-0.36	-1.40
Scale (cost vs safety)	0.16	0.49	+205%	-0.58	-2.56
Scale (time vs safety)	0.16	0.55	+248%	-0.77	-4.38

Table 4: Results for Danish data

TEST 1.1

	commuters	non-commuters
LL base	-2,404.02	-8,898.48
par	3	3
LL scale	-2,396.33	-8,877.16
choice tasks	8	8
LR-test	15.39	42.63
df	7	7
p-value	0.03	0.00

TEST 2.1

	commuters	non-commuters
LL combined	-2,368.71	-8,800.37
par	24	24
LR-test (comb vs unsc)	70.63	196.22
df	21	21
p-value	0.00	0.00
LR-test (comb vs sc)	55.24	153.59
df	14	14
p-value	0.00	0.00

TEST 1.2 (scale parameters)

	commuters		non-commuters	
Choice task	**est.**	**t-ratio (vs. 1)**	**est.**	**t-ratio (vs. 1)**
1	1	-	1	-
2	0.94	-0.10	0.35	-6.56
3	1.47	0.59	0.65	-2.03
4	2.06	0.97	0.55	-2.76
5	1.44	0.51	0.62	-2.20
6	2.35	0.99	0.59	-2.21
7	1.15	0.19	0.84	-0.70
8	1.21	0.27	0.61	-2.18

TEST 2.2 (LL per obs)

Task	commuters	non-commuters
1	-0.6241	-0.6079
2	-0.6716	-0.6807
3	-0.6549	-0.6629
4	-0.6458	-0.6712
5	-0.6592	-0.6622
6	-0.6323	-0.6657
7	-0.6622	-0.6472
8	-0.6659	-0.6608
correl (t-rat)	-0.33 (-0.87)	-0.27 (-0.68)
scaled model	-0.6596	-0.6631

TEST 1.3 (VTT)

	comm..	non-comm.
Unscaled est	55.52	33.59
Unscaled t-rat	12.20	10.90
Scaled est	53.44	32.74
Scaled t-rat	12.40	11.10
t-rat (diff)	-0.33	-0.20

Test 2.3 (relative heterogeneity)

	commuters			non-commuters		
	ASC	**cost**	**time**	**ASC**	**cost**	**time**
cv (scaled)	0.33	0.33	0.33	0.28	0.28	0.28
cv (individual)	2.09	0.36	0.31	1.30	0.28	0.37
rate of change	+543%	+9%	-5%	+363%	+1%	+33%
corr between task & est net of scale	-0.74	-0.69	-0.43	-0.87	-0.86	-0.75
t-rat	-2.70	-2.31	-1.17	-4.29	-4.19	-2.78

Table 5: Results for M4 commuter data

TEST 1.1

LL base	-2,854.28
par	7
LL scale	-2,846.23
choice tasks	16
LR-test	16.09
df	15
p-value	0.38

TEST 2.1

LL joint unscaled	-2,854.28
par	7
LR-test (comb vs unsc)	97.96
df	49
p-value	0.00
LR-test (comb vs sc)	81.86
df	34
p-value	0.00

TEST 1.3 (trade-offs against cost coefficient)

	free flow time	slowed down	travel time variability (AUD/hr)	toll
Unscaled est	13.21	16.74	1.11	1.68
Unscaled t-rat	7.7	10.1	11.2	1.8
Scaled est	13.14	16.728	1.12	1.65
Scaled t-rat	7.7	10.2	11.3	1.8
t-rat (diff)	-0.03	-0.01	0.02	-0.03

TEST 1.2 (scale parameters)

Choice task	est.	t-ratio
1	1	-
2	1.00	0.00
3	1.13	0.75
4	0.89	-0.81
5	1.02	0.13
6	1.27	1.33
7	1.11	0.70
8	1.26	1.38
9	1.17	0.92
10	1.22	1.18
11	1.16	0.88
12	1.20	1.05
13	1.24	1.34
14	1.19	1.02
15	1.21	1.14
16	1.34	1.75

Test 2.2 (trends in LL)

Task	per obs LL
1	-0.7890
2	-0.8021
3	-0.7511
4	-0.8262
5	-0.7949
6	-0.7181
7	-0.7456
8	-0.7253
9	-0.7329
10	-0.7068
11	-0.7430
12	-0.7376
13	-0.6722
14	-0.7068
15	-0.7021
16	-0.6831
correl (t-rat)	0.82 (5.33)
scaled model	-0.7506

Test 2.3 (relative heterogeneity)

	cv (scaled)	cv (individual)	rate of change	corr between task & est net of scale	t-rat
ASC1	0.10	1.01	+908%	-0.44	-1.83
ASC2	0.10	2.59	+2,473%	-0.33	-1.32
free flow time	0.10	0.18	+81%	0.26	0.99
slowed down time	0.10	0.24	+136%	0.35	1.39
travel time variability	0.10	0.13	+25%	0.10	0.36
cost	0.10	0.17	+65%	-0.50	-2.18
toll	0.10	0.70	+592%	0.39	1.58

Table 6: Results for M4 non-commuter data

TEST 1.1

LL base	-2,395.90
par	7
LL scale	-2,378.20
choice tasks	16
LR-test	35.40
df	15
p-value	0.00

TEST 2.1

LL joint unscaled	-2,395.90
par	7
LR-test (comb vs unsc)	146.60
df	49
p-value	0.00
LR-test (comb vs sc)	111.20
df	34
p-value	0.00

TEST 1.3 (trade-offs against cost coefficient)

	free flow time	slowed down	travel time variability (AUD/hr)	toll
Unscaled est	13.38	15.25	1.22	1.43
Unscaled t-rat	7.1	7.9	8.7	1.7
Scaled est	13.37	15.31	1.21	1.54
Scaled t-rat	7.3	8.1	9.1	1.9
t-rat (diff)	0.00	0.02	-0.02	0.09

TEST 1.2 (scale parameters)

Choice task	est.	t-ratio
1	1	-
2	1.27	1.32
3	0.95	-0.36
4	0.81	-1.60
5	1.54	2.05
6	0.96	-0.27
7	1.35	1.54
8	1.33	1.47
9	1.11	0.64
10	1.41	1.81
11	1.23	1.08
12	1.15	0.88
13	1.08	0.40
14	1.37	1.71
15	1.05	0.31
16	1.09	0.56

Test 2.2 (trends in LL)

Task	per obs LL
1	-0.7278
2	-0.7102
3	-0.7595
4	-0.8468
5	-0.6073
6	-0.7829
7	-0.6473
8	-0.6395
9	-0.7161
10	-0.6522
11	-0.6780
12	-0.6859
13	-0.7473
14	-0.6561
15	-0.7268
16	-0.7459
correl (t-rat)	0.16 (0.62)
scaled model	-0.7251

Test 2.3 (relative heterogeneity)

	cv (scaled)	cv (individual)	rate of change	corr between task & est net of scale	t-rat
ASC1	0.16	1.10	+571%	0.11	0.42
ASC2	0.16	1.23	+649%	-0.35	-1.40
free flow time	0.16	0.33	+100%	0.28	1.08
slowed down time	0.16	0.26	+57%	-0.26	-1.02
travel time variability	0.16	0.26	+60%	0.32	1.27
cost	0.16	0.17	+5%	0.12	0.44
toll	0.16	1.71	+944%	-0.12	-0.47

Table 7: Results for second Australian study

TEST 1.1

LL base	-2668.39
par	8
LL scale	-2644.62
choice tasks	16
LR-test	47.54
df	15
p-value	0.00

TEST 2.1

LL joint unscaled	-2,668.39
par	8
LR-test (comb vs unsc)	157.17
df	56
p-value	0.00
LR-test (comb vs sc)	109.63
df	41
p-value	0.00

TEST 1.3 (trade-offs against cost coefficient)

	free flow time (AUD/hr)	slowed down time (AUD/hr)	crawl time (AUD/hr)	travel time variability (AUD/hr)	toll ($/$)
Unscaled est	8.51	12.55	17.86	-3.33	0.89
Unscaled t-rat	3.70	5.70	5.70	1.10	6.60
Scaled est	8.45	13.03	17.95	-2.97	0.91
Scaled t-rat	3.60	5.60	5.60	1.00	6.30
T-rat (diff)	-0.02	0.15	0.02	0.09	0.09

TEST 1.2 (scale parameters)

Choice task	est.	t-ratio
1	1	-
2	0.96	-0.36
3	1.16	1.16
4	1.12	1.01
5	1.35	2.20
6	1.29	1.95
7	1.28	1.97
8	1.25	1.69
9	1.45	2.59
10	1.60	2.90
11	1.35	2.03
12	1.65	3.39
13	1.37	2.50
14	1.35	2.50
15	1.46	2.61
16	1.31	1.89

Test 2.2 (trends in LL)

Task	per obs LL
1	-0.6398
2	-0.6868
3	-0.6066
4	-0.5549
5	-0.5431
6	-0.5428
7	-0.5378
8	-0.5092
9	-0.4885
10	-0.4605
11	-0.5217
12	-0.3977
13	-0.5033
14	-0.5036
15	-0.5141
16	-0.5086
correl (t-rat)	0.74 (4.07)
scaled model	-0.5437

Test 2.3 (relative heterogeneity)

	cv (scaled)	cv (individual)	rate of change	corr between task & est net of scale	t-rat
ASC1	0.14	+0.86	+521%	0.12	0.45
ASC2	0.14	5.91	+4,151%	-0.10	-0.38
free flow time	0.14	0.52	+274%	0.35	1.39
slowed down time	0.14	0.35	+155%	-0.30	-1.16
crawl time	0.14	0.20	+47%	0.29	1.15
travel time variability	0.14	2.13	+1,428%	-0.18	-0.69
cost	0.14	0.34	+146%	0.43	1.80
toll	0.14	0.19	+36%	0.11	0.40

Table 8: MMNL results

	Atlanta sample 1	Atlanta sample 2	Atlanta sample 3	Danish commuters	Danish non-commuters
LL base	-7,705.63	-4,795.75	-4,410.44	-2,087.14	-7,231.21
LL scale	-7,694.40	-4,789.50	-4,406.11	-2,064.54	-7,211.95
LR-test	22.45	12.50	8.67	45.20	38.51
df	7	7	7	7	7
p-value	0.0021	0.0853	0.2775	0.0000	0.0000

Choice task	scale est.	t-ratio (vs. 1)	scale est.	t-ratio (vs. 1)	scale est.	t-ratio (vs. 1)	scale est.	t-ratio (vs. 1)	scale est.	t-ratio (vs. 1)
1	1	-	1	-	1	-	1	-	1	-
2	1.14	1.65	0.96	-0.55	1.02	0.24	0.75	-0.76	0.76	-1.28
3	1.15	1.64	1.02	0.24	1.08	0.90	1.94	1.38	1.24	0.77
4	1.18	1.98	0.90	-1.41	1.07	0.73	3.32	1.67	1.43	1.14
5	1.08	0.94	1.06	0.68	1.14	1.43	3.46	1.81	1.64	1.42
6	1.17	1.84	0.89	-1.48	1.18	1.85	3.17	1.90	1.65	1.43
7	1.11	1.21	0.96	-0.48	1.20	1.94	2.63	1.55	1.63	1.37
8	1.15	1.59	0.97	-0.32	1.17	1.67	2.91	1.77	1.97	1.56

	Fungibility	M4 commuters	M4 non-commuters	Second Australian
LL base	-2,786.28	-2,366.25	-1,978.79	-2,213.39
LL scale	-2,770.76	-2,347.10	-1,958.89	-2,178.50
LR-test	31.04	38.29	39.79	69.78
df	14	15	15	15
p-value	0.0055	0.0008	0.0005	0.0000

Choice task	scale est.	t-ratio (vs. 1)	scale est.	t-ratio (vs. 1)	scale est.	t-ratio (vs. 1)	scale est.	t-ratio (vs. 1)
1	1	-	1	-	1	-	1	-
2	0.50	-3.62	1.26	1.20	1.49	1.74	0.96	-0.22
3	0.70	-1.16	1.25	1.21	1.40	1.52	1.60	1.68
4	1.85	1.23	1.33	1.45	1.30	1.28	1.33	1.40
5	0.72	-1.36	1.39	1.66	1.79	2.30	1.79	2.05
6	0.62	-1.56	1.36	1.61	1.85	2.31	2.46	2.93
7	1.08	0.17	1.36	1.64	2.53	2.87	1.89	2.20
8	1.01	0.03	1.77	2.80	2.38	2.29	1.68	2.10
9	0.61	-1.64	1.55	2.08	1.63	1.97	2.59	2.56
10	0.72	-1.08	1.50	2.07	1.94	2.30	1.90	2.10
11	0.62	-1.56	1.49	1.91	1.78	2.31	2.17	2.46
12	0.92	-0.20	1.69	2.15	1.86	2.83	2.57	2.70
13	1.03	0.07	2.03	2.97	2.00	2.58	1.87	2.16
14	1.15	0.32	2.00	2.82	1.78	2.33	1.88	2.44
15	0.87	-0.38	1.64	2.25	1.63	2.24	2.48	3.01
16	-	-	1.78	2.60	1.68	2.30	1.71	2.18

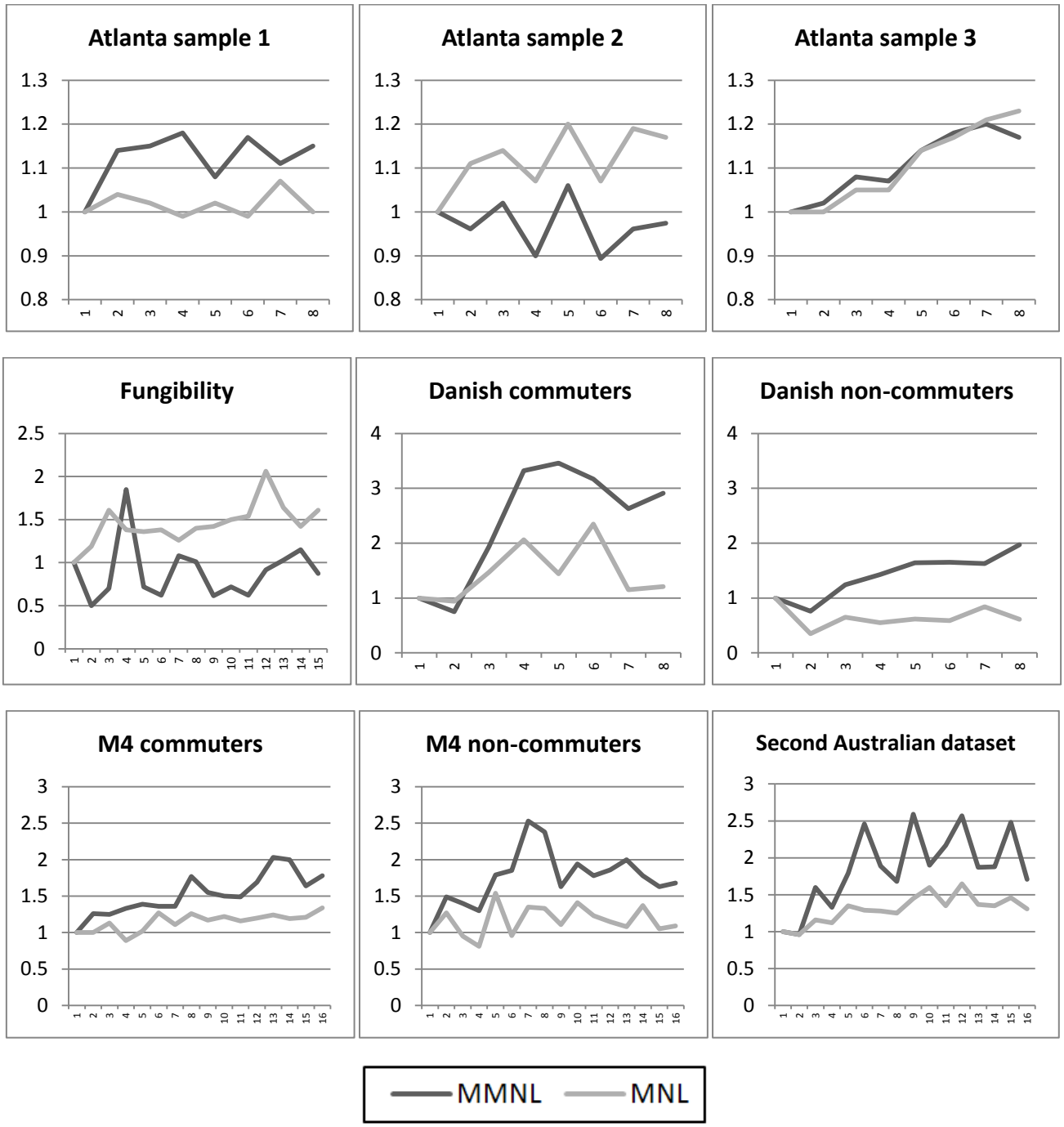

Figure 1: Summary results on choice task specific scale parameters
(choice task on x-axis, scale parameter on y-axis)

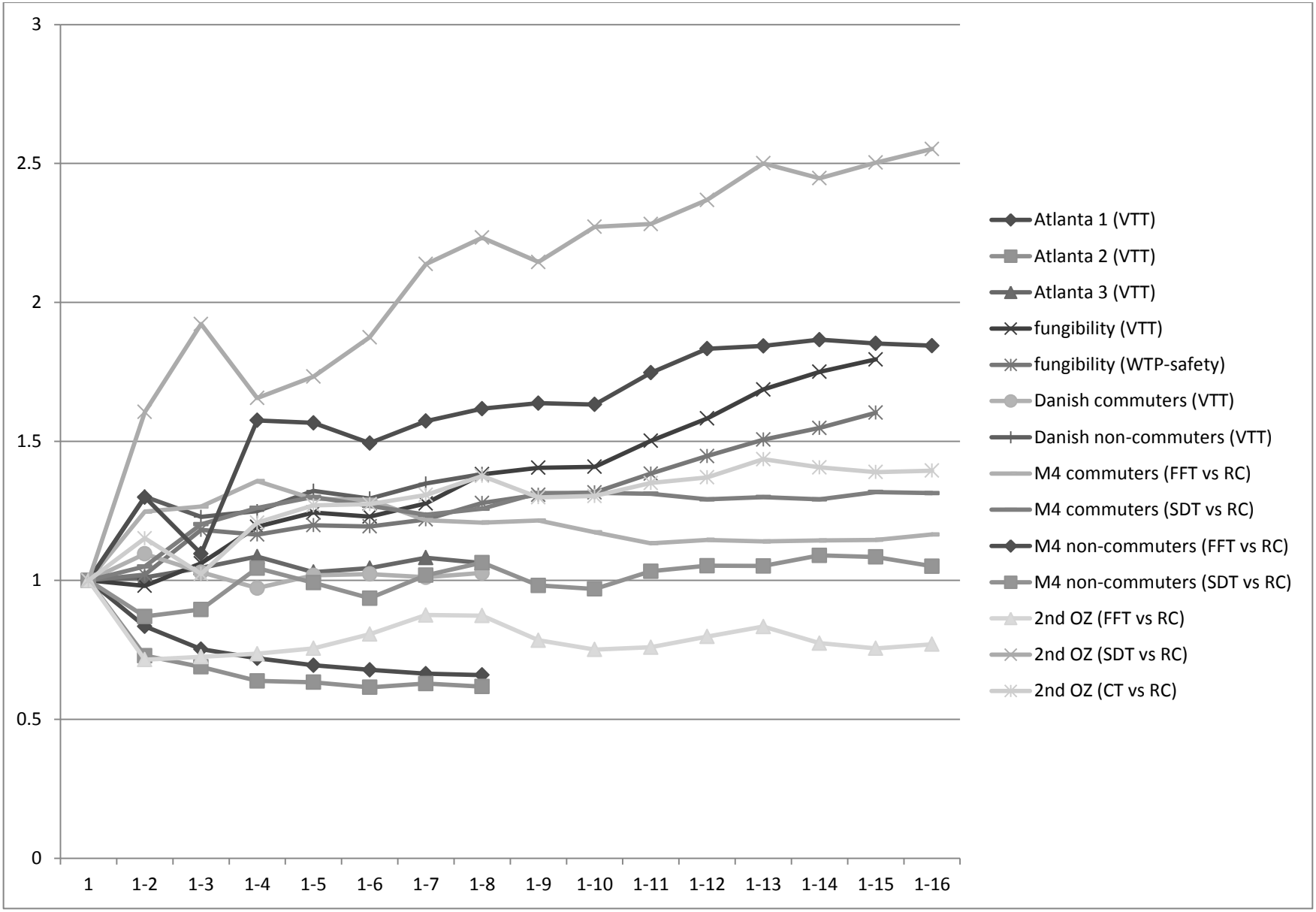

Figure 2: Trends in WTP measures
(range of choice tasks on x-axis, scale parameter on y-axis)