



UNIVERSITY OF WARSAW

Faculty of Economic Sciences

WORKING PAPERS

No. 8/2012 (74)

MIKOŁAJ CZAJKOWSKI
MAREK GIERGICZNY
WILLIAM H. GREENE

LEARNING AND FATIGUE EFFECTS REVISITED THE IMPACT OF ACCOUNTING FOR UNOBSERVABLE PREFERENCE AND SCALE HETEROGENEITY ON PERCEIVED ORDERING EFFECTS IN MULTIPLE CHOICE TASK DISCRETE CHOICE EXPERIMENTS

WARSAW 2012



UNIVERSITY OF WARSAW
Faculty of Economic Sciences

Learning and Fatigue Effects Revisited.

The Impact of Accounting for Unobservable Preference and Scale Heterogeneity on Perceived Ordering Effects in Multiple Choice Task Discrete Choice Experiments

Mikołaj Czajkowski
University of Warsaw
Faculty of Economic Sciences
e-mail: miq@wne.uw.edu.pl

Marek Giergiczny
University of Warsaw
Faculty of Economic Sciences
e-mail: mgiergiczny@wne.uw.edu.pl

William H. Greene
New York University
Stern School of Business, Department of Economics
e-mail: wgreene@stern.nyu.edu

Abstract

Using multiple choice tasks per respondent in discrete choice experiment studies increase the amount of available information. However, treating repeated choice data in the same way as cross-sectional data may lead to biased estimates. In particular, respondents' learning and fatigue may lead to changes in observed utility function preference (taste) parameters, as well as its error term variance (scale). Substantial body of empirical research offers mixed evidence in terms of whether (and which) of these ordering effects are observed. In this study we point to a significant component in explaining these differences – we show how accounting for unobservable preference and scale heterogeneity can influence the magnitude of observed ordering effects, especially if combined with too few choice tasks used for the analysis. We do this by utilizing the state-of-the-art modeling methods (H-MNL, S-MNL, H-RPL, G-MNL) which we modify to accommodate choice task specific scale parameter. In addition, we investigate possible bias resulting from not accounting for ordering effects. Our empirical study was based in the context of environmental protection – management changes in the protection of Polish forests.

Keywords:

ordering effects, learning, fatigue, preference and scale heterogeneity, forest management, recreation, biodiversity

JEL:

Q51, Q23, Q26, Q57

Acknowledgements:

The authors wish to thank Richard Carson whose valuable comments facilitated this study. Support of Polish Ministry of Science and Higher Education as well as Foundation for Polish Science is gratefully acknowledged.

Working Papers contain preliminary research results.

Please consider this when citing the paper.

Please contact the authors to give comments or to obtain revised version.

Any mistakes and the views expressed herein are solely those of the authors.

1. Introduction

One of the characteristics of a discrete choice experiment (DCE) study is that respondents are faced with more than one choice task. This may have some significant implications, as concerns for the stability of the preference structures become even more pronounced than in other stated preference methods. The accuracy of choices observed in a DCE study, as well as underlying decision strategies, may change as a respondent proceeds through multiple choice tasks. These phenomena are generally known as *ordering effects* and their manifestations are usually referred to in the literature, as (1) institutional learning, (2) value learning, (3) fatigue or boredom, and (4) choice set order dependence. These effects are not mutually exclusive and so many studies investigated them jointly.

Institutional learning relates to the fact that most respondents of a DCE have never participated in a stated preference survey before.¹ It is typically expected that institutional learning leads to an increase in the accuracy of responses – as respondents progress through choice tasks their responses are likely to become more accurate, at least until they complete their ‘burn-in’ number of choice tasks.² By more accurate (more deterministic) choices we mean lower variance of the utility function error term, i.e. higher utility function scale. This is what Swait and Adamowicz (2001a) call “smaller noise to signal ratio”.

Value learning can work in a similar way – as respondents complete more choice tasks they may discover or form their preferences, learn which attributes are the most significant, and what trade-offs

¹ A number of studies show that respondent’s consistency depends on the complexity of a choice task, such as the number of attributes, their levels, ranges, correlations, and the number of alternatives (e.g. Palma *et al.* 1994; Dellaert *et al.* 1999; Swait & Adamowicz 2001a; Caussade *et al.* 2005; Hensher 2006a, b; Day & Pinto 2010). This observation together with the hypothesis of information processing limitations supports institutional learning.

² However, as Swait and Adamowicz (2001b) note, if due to perceived high complexity of choice tasks or cumulative cognitive burden, respondents may simplify their choice strategies, adopting non-compensatory decision rules, one can also expect changes in estimated marginal utilities of the attributes.

they are willing to make. Eventually, this simplifies choice tasks for them and so is expected to reduce the error term variance. It is also possible that value learning can result in the change in respondents' preferences (tastes) associated with the attributes of choice between choice tasks. It is not clear how the effects of value and institutional learning could be told apart – it is usually assumed, however, that institutional learning takes place during the first few choice tasks only.

Fatigue or boredom can be expected to work in the opposite direction – increase the variance of the utility function error term. As a result it was postulated that since the effects of learning decrease, while effects of fatigue increase with choice task number, a U-shaped relationship between the weight for the unobservable part of utility function and the choice task number may hold.

Finally, ordering effects might occur as a result of path-dependency. In the case anchoring, framing, or acting strategically were happening, respondents' choices could become path-dependent. Theoretical analysis shows that these effects could influence decision rules in multiple ways (Carson & Groves 2007; Day *et al.* 2012).³

Empirical studies, discussed in detail in the next section, offer very contradictory conclusions with respect to the presence of learning and fatigue. These studies, however, differ significantly with respect to the methodology used to observe ordering effects – in particular with respect to the number of choice tasks, and whether unobservable preference and scale heterogeneity was accounted for. Our study aims at filling this gap.

In what follows, we show that controlling for the differences in methodology, in particular allowing for unobservable preference or scale heterogeneity, as well as using enough choice tasks per respondent play a significant role in whether (and which) ordering effects are observed. We argue that these differences allow us to explain some of the contradictory evidence presented by earlier studies.

³ Day *et al.* (2012) distinguish between starting point effect, acting strategically with full recall of the presented choice tasks, and acting strategically with imperfect recall, weighted towards more recent choice tasks. This distinction does not, however, matter for our analysis.

In addition, we propose the most flexible (so far) methods to control for ordering effects, and we investigate the possible implications of the bias resulting from not accounting for them.

The remainder of this paper is structured as follows. Section 2 reviews the results of earlier studies devoted to analyzing ordering effects and highlights the differences in methodologies used. In section 3 we present the new methods used in our study. Section 4 introduces details of empirical study designed specifically to investigate the issues of learning and fatigue – a large-scale national representative survey in the context of environmental protection – forest management in Poland. Section 5 reports the results. Our findings are discussed in section 6, which also offers conclusions relevant for future investigation of this research topic, and for the applications of the DCE methodology in general.

2. Ordering effects in the literature

Annex 1 provides a list of empirical studies devoted to investigating ordering effects. Existing evidence is mixed. While some studies report preference or scale changes resulting from institutional or value learning, others do not find statistically significant effects or observe respondents' fatigue. We note, however, that the studies differ substantially in the number of choice tasks used for the analysis – from as few as 4, to as many as 96 or 120 choice tasks per respondent (Brazell & Louviere 1997). A closer look at these studies, however, reveals tremendous variability not only with respect to the number of choice tasks, but also in the design, context, and the methods used to account for preference and scale differences.

There are many reasons why existing studies could observe different results in terms of whether, and how ordering effects manifest themselves. These include differences in the context of a study (e.g. familiar vs. unfamiliar goods, Oppewal *et al.* 2010; Day *et al.* 2012), overall choice complexity (Swait

& Adamowicz 2001a; Caussade *et al.* 2005; Chung *et al.* 2010), cumulative cognitive burden (Swait & Adamowicz 2001b), and administration mode (e.g. mail vs. web survey, Savage & Waldman 2008).

Anchoring, framing or acting strategically may also lead to changes in respondents' choices. Therefore, it is necessary to control for these sources of ordering effects, if one wishes to isolate the effects of learning and fatigue. Many studies did not effectively take this into account. For instance using a design in which a sequence of choice tasks is repeated at the end of the series allows one to observe possible deviations in choice patterns; researchers are then able to observe changes in observed marginal utilities when comparing implicit prices between the subsets of choices, or differences in variance of the error term estimated for choices made at different times.⁴ Utilizing such a design, however, makes it impossible to control for path-dependent ordering effects, caused by anchoring, framing, or acting strategically, and they might be taken as evidence of learning or fatigue, or mask their influence.

One way to minimize the effects of anchoring, framing or acting strategically is to use a counterbalanced design, i.e. present each respondent with a different order of choice tasks, so that a potential effect of e.g. starting point (anchoring) is canceled out for the sample. Counterbalancing of a DCE design plays an important role in retrieving underlying dynamics of ordering effects (Keppel & Wickens 2004). Liechty *et al.* (2005) show theoretically that without counterbalancing, static models fail to capture average preference changes and dynamic models can give relatively poor estimates. Some examples of counterbalancing include cyclical design, in which choice tasks are rotated for different respondents (i.e. choice sets 1,2,3... T for the 1'st respondent, 2,3,..., T ,1 for the 2'nd respondent, and so on), and Latin square design, in which the cycling is reversed (Street & Street 1987). It seems better, however, to randomize the order of choice tasks for each respondent, in which case the sequence not only starts in a different point, but also is shuffled, i.e. choice sets do not appear in the same order.

⁴ Alternative treatments focused on observing time the respondents took to complete a choice task (Haaijer *et al.* 2000; Rose & Black 2006).

We note that the empirical studies listed in Annex 1 used a wide range of econometric models that, as we will show, can influence the observed result. Most studies repeated a choice task (or a sequence of choice tasks) and observed differences in implicit prices or scale, included interactions of choice attributes or scale with the choice task number, included choice task number as a covariate of scale, or as an explanatory variable of latent class membership (latent class model). The most common ways used to control for scale differences were: the procedure proposed by Ben-Akiva and Morikawa (1990), in which observations from separate (groups of) choice tasks are used simultaneously to maximize joint likelihood function, the Bradley and Daly (1992); (1994) one-step estimation approach of Ben-Akiva and Morikawa, which can be implemented using a nested logit (the *logit-based scaling approach*), the *Swait-Louviere procedure* (Swait & Louviere 1993) – a sequential scaling approach, and the Heteroskedastic Multinomial Logit (H-MNL) model with choice task specific covariates of scale (Hensher *et al.* 1998; Dellaert *et al.* 1999; Swait & Adamowicz 2001a).

More importantly, these studies differ substantially in the way unobservable preference or scale heterogeneity was treated. In the recent years, much research effort has been devoted to the issue of consumers' unobservable preference heterogeneity (e.g. McFadden & Train 2000). Another research stream aiming at accounting for consumers' heterogeneity has focused on modeling unobservable scale differences (e.g. Louviere *et al.* 2002). In what follows, we show that allowing for unobservable preference or scale heterogeneity may significantly influence the strength of observed learning and fatigue effects. Therefore, using a different model specification is an important factor that allows one to explain the mixed evidence presented by empirical studies investigating ordering effects.

3. Methods for modeling ordering effects

In this section we review the methods for accounting for unobservable preference and scale heterogeneity in discrete choice models. We later apply these methods to investigate how different model specifications may influence ordering effects.

3.1 Heteroscedastic Multinomial Logit Model (H-MNL)

The modeling of discrete choice data is built on random utility theory developed most notably by McFadden (1986). It assumes that the utility associated with any state (choice) can be divided into a sum of contributions that can be observed by a researcher, and a component that cannot, and hence is assumed random. This formulation of utility function and choice-specific alternatives leads to the multinomial logit model that allows using observed choices of an individual to compare their utility levels associated with the choice alternatives.

Formalizing, let individual i choose among J alternatives, each characterized by a vector of observed attributes \mathbf{x}_{ij} . The utility associated with alternative j is given by:

$$U_i(\text{Alternative} = j) = U_{ij} = \boldsymbol{\beta}' \mathbf{x}_{ij} + \varepsilon_{ij} \quad (1)$$

where $\boldsymbol{\beta}$ is a parameter vector of marginal utilities of the attributes. By introducing the error term it is assumed that utility levels are random variables, as it is otherwise impossible to explain why apparently equal individuals (equal in all attributes which can be observed) may choose different options.

Random utility theory is transformed into different classes of choice models by making different assumptions about random term. In order for the random component to represent the necessary amount

of randomness into respondents' choices, its variance needs to be sufficiently large or, since utility function has no scale, assumptions with respect to the random term variance may be expressed by scaling the utility function in the following way:

$$U_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \varepsilon_{ij} / \sigma, \quad (2)$$

where the random component of the utility function is conveniently assumed to be independently and identically (iid) distributed across individuals and alternatives – Extreme Value Type 1 distribution. The scale coefficient σ and $\boldsymbol{\beta}$ cannot *both* be identified. The multinomial logit model (MNL) is derived, with the following closed-form expression of the probability of choosing alternative j from a set of J available alternatives:

$$P(j|J) = \frac{\exp(\sigma \boldsymbol{\beta}'\mathbf{x}_{ij})}{\sum_{k=1}^J \exp(\sigma \boldsymbol{\beta}'\mathbf{x}_{ik})}. \quad (3)$$

The heteroscedastic MNL model allows the scale for some observations to systematically differ from the others. The utility specification of the H-MNL, with covariates of scale entering linearly (Dellaert *et al.* 1999) is:

$$U_{ij} = \sigma(1 + \boldsymbol{\theta}'\mathbf{k}_i) \boldsymbol{\beta}'\mathbf{x}_{ij} + \varepsilon_{ij}, \quad (4)$$

while by assuming an exponential formulation for the multiplicative scale (Swait & Adamowicz 2001a) it is possible to drop the 1 and there is no need to assume the scale is strictly positive:

$$U_{ij} = \sigma \exp(\boldsymbol{\theta}'\mathbf{k}_i) \boldsymbol{\beta}'\mathbf{x}_{ij} + \varepsilon_{ij}. \quad (5)$$

In both cases, the ‘effective’ scale is a function of \mathbf{k}_i – vector of respondent- or choice-specific and observable variables. The scale is still normalized, but with respect to the reference group and so it can differ for selected observations (e.g. choice tasks occurring later in the sequence).

3.2 Scale heterogeneity model (S-MNL)

The MNL model implausibly assumes that the random term is independent and identical for all choices, i.e. the scale coefficient is the same for every respondent, every choice task and every alternative. This results in assuming that every respondent makes his choices with the same degree of randomness. One way to relax this assumption is allowing the (unobservable) scale coefficient to be individual-specific, through making it a random variable following a particular (usually log-normal) distribution. The new utility specification becomes:

$$U_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \varepsilon_{ij} / \sigma_i, \quad (6)$$

where $\sigma_i \sim LN(1, \tau)$ or $\sigma_i = \exp(\bar{\sigma} + \tau\varepsilon_{0i})$ with $\varepsilon_{0i} \sim N(0, 1)$. Note that the scale coefficient is now respondent-specific. Since it is still convenient to normalize scale to 1, we want $E\sigma_i = \exp(\sigma + \tau^2/2)$. This may be achieved by assuming $\bar{\sigma} = -\tau^2/2$. This way the scale is no longer fixed; instead it is assumed to follow a lognormal distribution, with the new parameter τ reflecting the level of scale heterogeneity in the sample. The resulting model is a scale heterogeneity model (S-MNL, Fiebig *et al.* 2010).

In order to use this model to observe scale changes between choice tasks, we introduced additional explanatory variables of scale \mathbf{k} , such that $\sigma_i = \exp(\bar{\sigma} + \tau\varepsilon_{0i} + \boldsymbol{\theta}'\mathbf{k}_{it})$. These may be e.g. choice task specific variables, provided that one of the choice tasks is used as a reference level.

3.3 Heteroscedastic Random Parameters Logit model (H-RPL)

Another implausible assumption of the MNL model is that all respondents have the same preferences (and so the same coefficients in their utility functions, $\boldsymbol{\beta}$). The state-of-practice methods of relaxing

these assumptions, i.e. allowing for some level of (unobservable) preference heterogeneity and possibly correlations between the alternatives and choice tasks, include the Random Parameters Model (RPL, McFadden & Train 2000; Hensher & Greene 2003).

In RPL the utility function becomes:

$$U_{ij} = \sigma \beta_i' \mathbf{x}_{ij} + \varepsilon_{ij}. \quad (7)$$

Note that parameters of utility functions are now respondent-specific. It is assumed that they follow distributions specified by a modeller: $\beta_i \sim f(\mathbf{b} + \Delta' \mathbf{z}_i, \Sigma_i)$, with means \mathbf{b} and variance-covariance matrix Σ . In addition, it is possible to make means and variances of the distributions a function of observable respondent or choice-specific characteristics \mathbf{z} .

In this paper we refer to the Heteroscedastic RPL (H-RPL) as the RPL model in which scale is allowed to systematically differ for some observations. This is a natural extension of a H-MNL model, although to our knowledge it has never been used before.

The utility specification in H-RPL model becomes:

$$U_{ij} = \sigma \exp(\theta' \mathbf{k}_{it}) \beta_i' \mathbf{x}_{ij} + \omega_{ij}, \quad (8)$$

where the scale is a function of observable explanatory variables \mathbf{k} – in our case choice task specific, but generally any observable variables, as long as the scale may be normalized for some reference level observations.

3.4 Generalized Multinomial Logit Model (G-MNL)

A method which allows to control for both preference and scale heterogeneity of respondents at the same time is the Generalized Multinomial Logit Model (G-MNL, Fiebig *et al.* 2010). In this model, the utility function takes the form:

$$U_{ij} = [\sigma_i \mathbf{b} + \gamma \mathbf{v}_i + (1 - \gamma) \sigma_i \mathbf{v}_i]' \mathbf{x}_{ij} + \omega_{ij}. \quad (9)$$

Similarly to the RPL model, the coefficients in the utility function are individual-specific (\mathbf{b} represents the population means of the parameters, while \mathbf{v} – individual-specific deviations from these means). Unlike in the RPL, however, the scale coefficient is also individual-specific. The new coefficient $\gamma \in [0, 1]$ ⁵ controls how the variance of residual taste heterogeneity varies with scale. If $\gamma = 0$ the individual coefficients become $\beta_i = \sigma_i \mathbf{b} + \mathbf{u}_i$, while if $\gamma = 1$ they are $\beta_i = \sigma_i (\mathbf{b} + \mathbf{u}_i)$. These are the two extreme cases of scaling (or not scaling) residual taste heterogeneity in the G-MNL model (type I and type II respectively), however, all intermittent solutions are possible.

In estimation, the individual scale is normalized in the same way as in the S-MNL model. In addition, in this paper we extend the G-MNL model by making the individual scale parameter a function of observable (choice task specific) characteristics \mathbf{k} :

$$\sigma_i = \exp(\bar{\sigma} + \tau \varepsilon_{0i} + \boldsymbol{\theta}' \mathbf{k}_{it}). \quad (10)$$

This way we are able to investigate learning and fatigue effects while allowing for both preference and scale heterogeneity across respondents.

4. Empirical study

Our empirical study was based in the context of environmental protection – management changes in the protection of Polish forests. We were interested in the attributes of the Polish forests that are the most significant for the general public in terms of recreation and biodiversity conservation. Through

⁵ To assure $\gamma \in [0, 1]$ it is usually modeled as $\gamma = \frac{\exp(\gamma^*)}{1 + \exp(\gamma^*)}$, and it is γ^* that is estimated.

the extensive qualitative studies we discovered that the forest attributes that Poles would like to see changed the most were: (1) protection of the most ecologically valuable forests, (2) less litter in forests, and (3) an increasing the amount of recreational infrastructure. These were the attributes that we used for the hypothetical scenario of our DCE study.

Of the 90 000 km² Polish forests about 3% are forests which are the most ecologically valuable in terms of having many of the characteristics of natural forests, such as age and structure of trees, the presence of natural environmental processes, large amounts of dead wood, rare species of fauna and flora and high biodiversity in general (see Annex 2a for illustration). About 50% of these forests are currently properly protected, usually in the form of national parks and nature reserves. The rest is under much human pressure and often is treated as regular economic forests. Annex 3 provides a map of locations and areas of the most ecologically valuable forests in Poland. Therefore, the first attribute in our CE scenario was the area change of ecologically valuable forests that could be protected. The possible levels of this attribute were:



Status quo

Passive protection of **50%** of the most ecologically valuable forests
(1,5% of all the forests)



Partial improvement

Passive protection of **75%** of the most ecologically valuable forests
(2,25% of all the forests, 50% increase)



Substantial improvement

Passive protection of **100%** of the most ecologically valuable forests
(3% of all the forests, 100% increase)

The second attribute used in the final study was the amount of litter that was present in the forest. This could be left in forests by tourists or as illegal trash-dump sites (see Annex 2b for illustration). Litter obviously decreases recreational value of a forest, may leak dangerous substances, and constitutes a

hazard for animal life and health. In our hypothetical scenario it was proposed to reduce the amount of litter by 50% or by 90%, though tougher law enforcement and increasing forest cleaning services. The available levels of this attribute were:



Status quo

No change in the amount of litter in the forests



Partial improvement

Decrease the amount of litter in the forests by half
(50% reduction)



Substantial improvement

Litter found in the forests only occasionally
(90% reduction)

Qualitative pretesting also showed that for the recreational value of forests it was important that enough tourist infrastructure was available. This could include local roads allowing easier access to a forest, parking places, paths and trails for tourists, organized resting areas (e.g. picnic sites) or toilets. Our scenario proposed and described two levels of increased amount and quality of infrastructure. It was explained that such infrastructure would be built only where necessary and only in a way that does not influence the environment. In short, these were:



Status quo

No change in tourist infrastructure



Partial improvement

Appropriate tourist infrastructure in **a half more** forests
(50% increase)



Substantial improvement













Appropriate tourist infrastructure available in **twice more** forests
(100% increase)

The last attribute was monetary – additional annual cost per household, in the form of increased income taxes.

The final survey was conducted on a representative sample of 1001 Poles. We hired a professional polling agency that collected the questionnaires using high-quality, face-to-face computer-assisted surveying techniques. The choice sets utilized in our study were prepared using Bayesian d-efficient design optimized for the RPL model (Sándor & Wedel 2001; Ferrini & Scarpa 2007; Bliemer *et al.* 2008; Scarpa & Rose 2008). To obtain initial estimates (priors) and to verify the qualitative properties of the questionnaire itself we conducted a pilot study on a sample of 50 respondents.

Each respondent was faced with 26 choice tasks, each consisting of 4 alternatives. Each alternative was described with the 4 attributes specified above. An example of a choice card shown to respondents is given in Figure 1.

Figure 1. Example of a choice card

	I wariant	IV wariant	II wariant	III wariant
Ochrona najcenniejszych lasów	 Bez zmian Ochrona ścisła 50% najcenniejszych lasów (1,5% wszystkich lasów)	 Bez zmian Ochrona ścisła 50% najcenniejszych lasów (1,5% wszystkich lasów)	 Bez zmian Ochrona ścisła 50% najcenniejszych lasów (1,5% wszystkich lasów)	 Znaczna poprawa Ochrona ścisła 100% najcenniejszych lasów (3% wszystkich lasów, wzrost o 100%)
Śmieci w lasach	 Bez zmian Ilość śmieci w lasach taka jak obecnie	 Częściowa poprawa O połowę mniej śmieci w lasach (spadek o 50%)	 Bez zmian Ilość śmieci w lasach taka jak obecnie	 Częściowa poprawa O połowę mniej śmieci w lasach (spadek o 50%)
Infrastruktura	 Bez zmian Stan infrastruktury taki jak obecnie	 Bez zmian Stan infrastruktury taki jak obecnie	 Znaczna poprawa Odpowiednia infrastruktura na obszarze lasów dwa razy większym niż obecnie (wzrost o 100%)	 Częściowa poprawa Odpowiednia infrastruktura na obszarze lasów o połowę większym niż obecnie (wzrost o 50%)
Koszt	0 zł	10 zł	25 zł	100 zł

A significant contribution of our empirical work comes from the fact that our design was counterbalanced. This allowed us to control for the ordering effects occurring as a result of anchoring, framing or acting strategically. We achieved this by randomizing the order of 26 choice tasks presented to each respondent. In addition, we randomized the order of the 3 non-status-quo alternatives for each choice task and each respondent. This treatment allows us to focus on the effects of learning and fatigue, while potential effect of starting point (anchoring) is canceled out.

5. Results

5.1. Preference (taste) dynamics

We start the analysis by testing if there are systematic changes in respondents' preferences associated with the attributes of choice. We do this by estimating a separate MNL model for each choice task, as this approach allows utility function coefficients and scale to be fully choice task specific. However, since the parameters of each choice task specific model are confounded with a different scale, it is not possible to directly compare their values between the models. Instead, we calculated implicit prices for each of the choice attributes and each choice task – this way model specific scale is canceled out and we are able to observe potential dynamics of implicit prices.

Even though in the following analysis we applied other, more sophisticated models, we argue that for the task of analyzing dynamics of implicit prices based on choice task specific models using MNL model is appropriate. There are two main reasons for this. First of all, estimating choice task specific models that allow for scale heterogeneity (S-MNL), preference heterogeneity (RPL), or both (G-MNL) proves difficult, as with only a single observation for each respondent it is difficult to distinguish between random heterogeneity and the IID extreme value term in the model (Ruud 1996; Revelt &

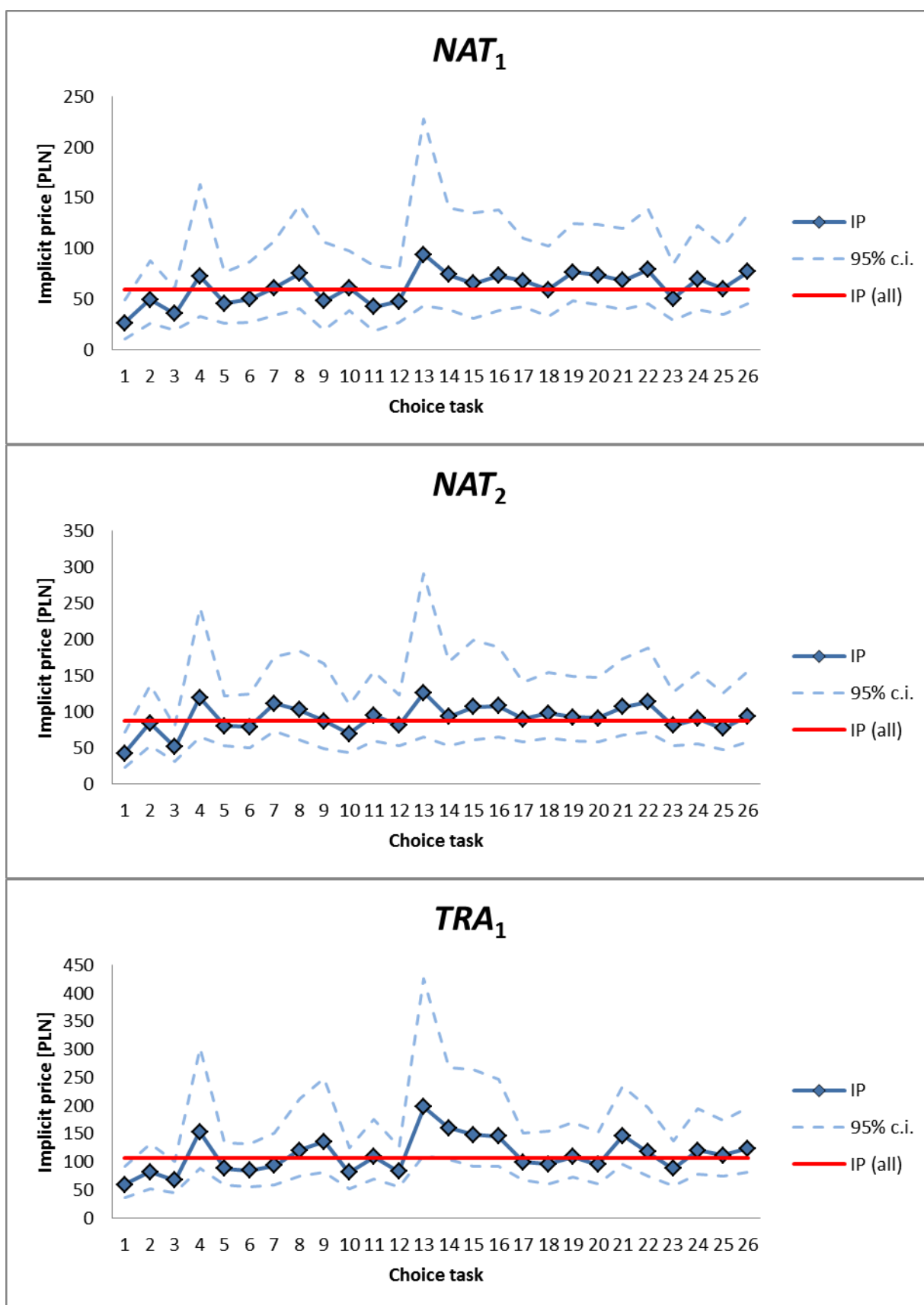
Train 1998; Fosgerau & Nielsen 2010; Hess & Train 2011). Secondly, standard errors associated with implicit prices derived from MNL model are the narrowest and so allow for the most conservative comparison criterions. The models allowing for heterogeneity with respect to attribute coefficients result in much wider confidence intervals of implicit prices, as they result from not only standard errors associated with coefficients, but also from standard deviations of empirical distributions of random coefficients (as well as standard errors associated with their means and standard deviations). Therefore, using confidence intervals derived from a simple MNL model seems the most conservative way of the analysis of statistical differences between choice task specific implicit prices.

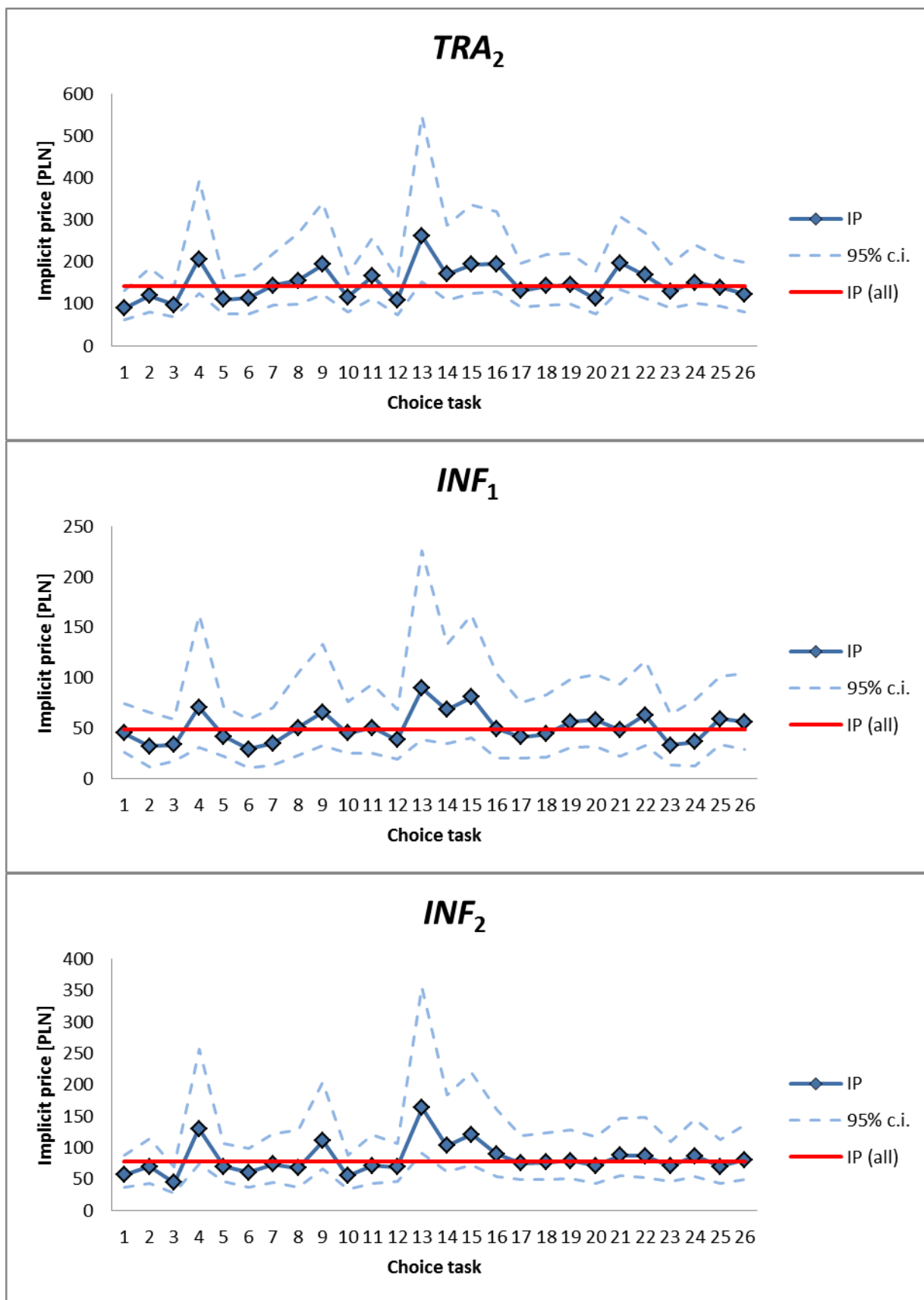
The 6 panels of Figure 2 present the dynamics of choice task specific implicit prices associated with the choice attributes (IP) along with 95% confidence intervals, and implicit prices derived from MNL models estimated for all choice tasks (IP-all). The qualitative attributes were dummy coded with status quo as a reference level, and so the presented implicit prices are:

- NAT_1 , NAT_2 – partial (50%) and substantial (100%) improvement in the area of passively protected ecologically valuable forests,
- TRA_1 , TRA_2 – partial (50%) and substantial (90%) reduction of litter in the forests,
- INF_1 , INF_2 – partial (50%) and substantial (100%) increase of forests with tourist infrastructure present.

The analysis of implicit prices dynamics shows that there is a considerable amount of variation between choice tasks. However, no systematic patterns are visible and implicit prices are not statistically different between choice tasks, as indicated by overlapping confidence intervals. This finding is supported by a similar analysis based on the results of other models mentioned above – taking preference and scale heterogeneity into account results in substantially wider confidence intervals of implicit prices and even less significant differences between choice task specific implicit prices.

Figure 2. Dynamics of choice task specific implicit prices





In conclusion, we do not find significant preference changes between choice tasks. This result is in line with many earlier empirical studies which do not observe statistically significant differences in preferences (tastes) between choice tasks (see Annex 1 for details). We note, however, that the mixed evidence provided by some of these studies is related to problems with identifying intra-responder taste heterogeneity, as discussed by Hess and Train (2011). In particular low statistical power of the tests may result in the taste dynamics rarely being statistically significant.

Based on the findings reported in this section, in what follows, we assumed that individual preference (taste) parameters are constant throughout the choice experiment and focused on analyzing scale dynamics and potential bias resulting from not accounting for them.

5.2. Scale dynamics

In order to investigate scale dynamics under different methodological assumptions, we estimated H-MNL, S-MNL, H-RPL and G-MNL models (the last two with and without allowing for correlations of random parameters)⁶ in which dummy coded choice task numbers ($CS_2 - CS_{26}$) entered as explanatory variables of scale (the first choice task was used as a reference level). All random parameters associated with the choice attributes were assumed to be normally distributed. Where applicable, we accounted for the panel structure of our dataset (since each respondent faced 26 choice tasks) by introducing random effects type of treatment – additional random term for all observations from the same individual. All models were estimated using 26,026 choice observations and 1000 random draws.

The results are presented in Table 1. Panel 1 presents coefficients associated with choice attributes or their means for models that assume they are random. In addition to the variable names used earlier, we

⁶ In what follows we used suffix ‘_d’ for RPL and G-MNL models in which only elements on the diagonal of the Cholesky matrix are estimated (no correlations are allowed), and no suffix otherwise.

provide estimates for FEE – the cost coefficient – and SQ – alternative specific constant associated with the status quo (no action) alternative. In Panel 2 we present standard deviations of normally distributed parameters. Panel 3 contains coefficients of parameters associated with scale, i.e. τ and γ (where applicable), and 25 choice task specific covariates of scale. The estimates of below-diagonal elements of Cholesky matrices were omitted for brevity but are available from the authors upon request.

We begin the analysis by noting that all explanatory variables turn out to be significant determinants of choices and are of expected sign. The statistical significance of the coefficients associated with the standard deviations of the random parameters distributions indicates that they are significantly different from zero, and hence that the variables should indeed be modeled as random. This is strong evidence of unobservable preference heterogeneity. On the other hand, in S-MNL and both G-MNL models the τ coefficient representing dispersion of individual scale coefficients is significantly different from 0 that indicates considerable (unobservable) heterogeneity in individual scale coefficients. Finally, as expected, models that allow for unobservable preference and scale heterogeneity generally perform better than models that do not; the same holds for the models that allow for correlations between random parameters vs. the ones that do not.

In order to facilitate the examination of scale dynamics Figure 3 presents choice task specific scale parameters of each model, along with their 95% confidence intervals. A general pattern becomes apparent – irrespectively of the model used, the scale appears to increase until about 8th choice task and then stabilizes – even though some degree of variability is present, the scale does not seem to decrease. This corresponds to respondents’ learning – their choices become more deterministic – and at the same time we do not observe effects of fatigue that would lead to increased randomness in later choice tasks of our study.

Secondly, we note that the more sophisticated model is used, in terms of allowing for unobservable preference and scale heterogeneity, the higher an increase in choice task specific scale parameter. The magnitude of this increase grows from allowing for scale heterogeneity only (S-MNL) through

allowing for preference heterogeneity (H-RPL) to allowing for both (G-MNL). Interestingly, the choice task specific scale coefficients for H-RPL_d model (without correlations) are lower than those for H-RPL (with correlations). This is in line with our general findings, as allowing for random parameters correlations introduces some degree of specific scale heterogeneity.

Accounting for preference and scale heterogeneity may change the conclusions on whether effects of learning or fatigue are observed in a study, especially if relatively few choice tasks are used. In our case, the results for H-MNL show that scale does not become statistically different from 1 until choice task 10. This is not the case in S-MNL, H-RPL or G-MNL where the scale increase is faster. Therefore, depending on what model specification was used in the analysis, different conclusions could be drawn. This is one of the main conclusions resulting from our analysis.

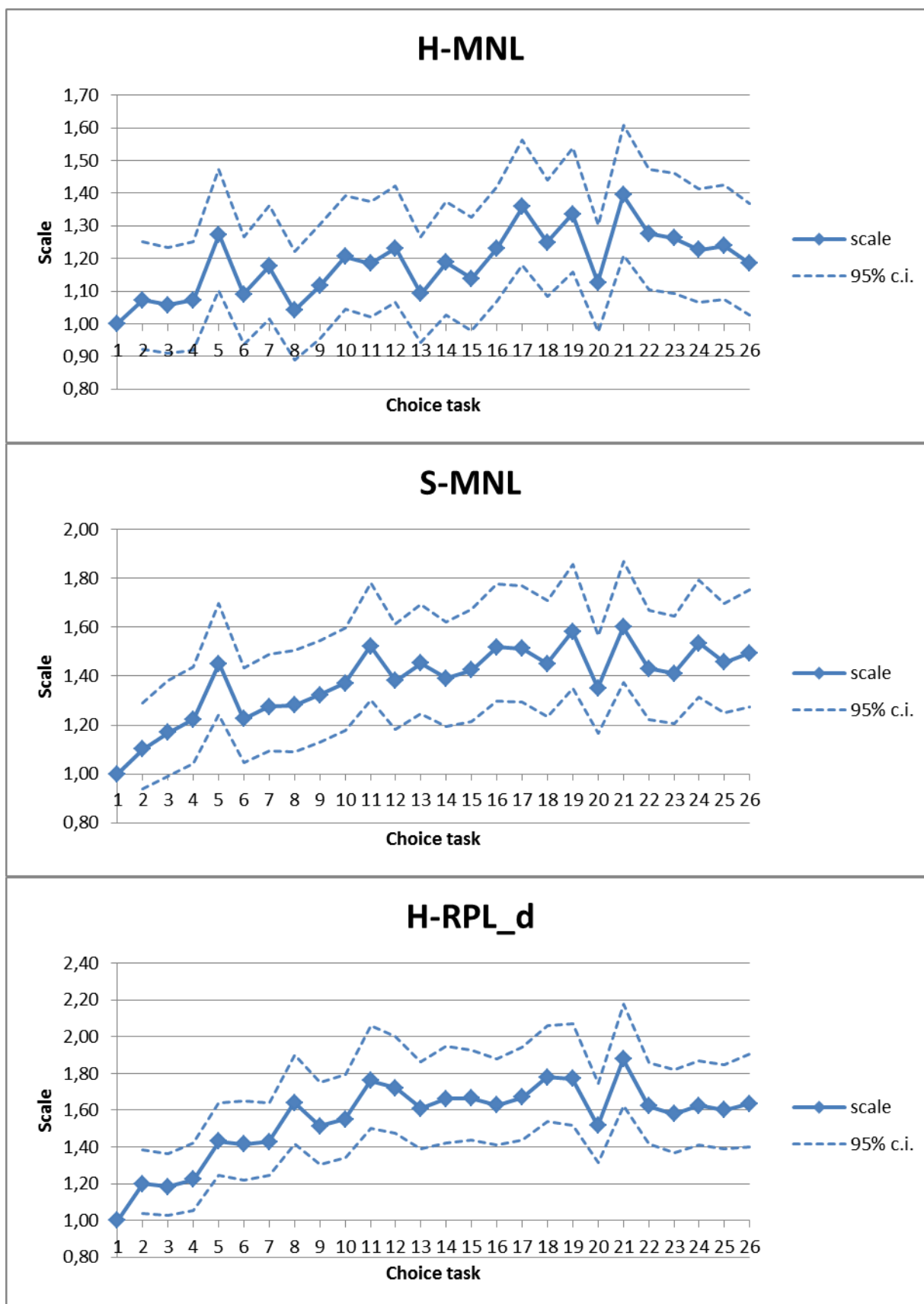
Finally, we note that even for studies that do allow for preference and scale heterogeneity, using too few choice tasks results in confidence intervals of choice task specific scale being wider. We illustrate this in Figure 4, where choice task specific scale and its 95% confidence intervals were presented for only the first 4 choice tasks. Using the models estimated on data from only the first 4 or all 26 choice tasks influences the ability to identify preference and scale heterogeneity, and as a result influences the confidence intervals of choice task specific scale. This effect is not present for H-MNL, since in this case no unobservable preference or scale heterogeneity is allowed. However, for other models, using too few choice tasks could lead to wider confidence intervals, and as a result, different conclusions in terms of whether learning or fatigue effects are present.

Table 1. Scale dynamics under different model specifications (standard errors given in parentheses)

	H-MNL	S-MNL	H-RPL_d	H-RPL	G-MNL_d	G-MNL
(1) Location parameters						
NAT_1	0.6761 (0.0422)	2.0561 (0.4780)	0.8703 (0.0483)	0.9614 (0.0561)	1.1756 (0.0719)	1.2133 (0.0789)
NAT_2	0.9944 (0.0584)	2.8274 (0.6586)	1.2806 (0.0687)	1.2891 (0.0770)	1.6752 (0.1009)	1.6342 (0.1090)
TRA_1	1.2154 (0.0695)	2.6301 (0.6088)	1.2203 (0.0622)	1.1409 (0.0636)	1.5324 (0.0887)	1.4382 (0.0891)
TRA_2	1.6229 (0.0912)	3.9680 (0.9212)	1.7811 (0.0921)	1.6978 (0.0937)	2.2691 (0.1327)	2.0891 (0.1306)
INF_1	0.5488 (0.0365)	1.3039 (0.3087)	0.6032 (0.0341)	0.5981 (0.0402)	0.7822 (0.0498)	0.7350 (0.0541)
INF_2	0.8867 (0.0520)	1.9245 (0.4481)	0.8842 (0.0489)	0.8364 (0.0503)	1.1119 (0.0695)	1.0388 (0.0705)
FEE	-1.1333 (0.0683)	-3.7883 (0.8865)	-2.8402 (0.1568)	-3.0452 (0.1684)	-3.6242 (0.2227)	-4.1801 (0.2557)
SQ	1.7000 (0.0979)	-0.4687 (0.1414)	-1.0375 (0.1085)	-1.5779 (0.1237)	-2.7315 (0.2208)	-1.5874 (0.1445)
(2) Standard deviations						
NAT_1	—	—	0.3840 (0.0320)	0.7905 (0.0456)	0.4852 (0.0368)	2.1058 (0.1435)
NAT_2	—	—	0.7118 (0.0413)	0.1374 (0.0299)	0.7149 (0.0483)	4.4034 (0.2942)
TRA_1	—	—	0.3304 (0.0317)	0.5612 (0.0444)	0.0804 (0.0382)	0.8569 (0.0608)
TRA_2	—	—	0.6989 (0.0417)	0.3903 (0.0313)	0.6266 (0.0431)	0.1091 (0.0353)
INF_1	—	—	0.1224 (0.0388)	0.2691 (0.0357)	0.1373 (0.0410)	0.4619 (0.0455)
INF_2	—	—	0.4124 (0.0303)	0.1533 (0.0378)	0.4447 (0.0348)	0.5217 (0.0409)
FEE	—	—	2.7742 (0.1500)	1.6153 (0.1113)	3.1858 (0.1920)	0.1146 (0.0564)
SQ	—	—	4.7041 (0.2637)	3.5832 (0.2223)	6.6775 (0.4273)	0.2286 (0.0409)
(3) Scale parameters						
τ	—	1.9004 (0.0776)	—	—	0.6477 (0.0202)	0.6876 (0.0235)
γ^*	—	—	—	—	-1.5218 (0.1385)	-1.6900 (0.1546)

CS_2	0.0702 (0.0781)	0.0962 (0.0807)	0.1811 (0.0740)	0.1987 (0.0769)	0.1580 (0.0777)	0.2237 (0.0802)
CS_3	0.0565 (0.0776)	0.1563 (0.0845)	0.1673 (0.0725)	0.2447 (0.0747)	0.2033 (0.0790)	0.2331 (0.0786)
CS_4	0.0699 (0.0782)	0.2017 (0.0826)	0.2023 (0.0761)	0.2970 (0.0801)	0.2695 (0.0817)	0.3338 (0.0818)
CS_5	0.2414 (0.0739)	0.3721 (0.0800)	0.3577 (0.0695)	0.4368 (0.0724)	0.4104 (0.0734)	0.4766 (0.0732)
CS_6	0.0861 (0.0764)	0.2028 (0.0801)	0.3478 (0.0774)	0.4121 (0.0775)	0.3755 (0.0805)	0.4542 (0.0823)
CS_7	0.1624 (0.0749)	0.2435 (0.0791)	0.3562 (0.0703)	0.4508 (0.0750)	0.4226 (0.0757)	0.4974 (0.0779)
CS_8	0.0407 (0.0812)	0.2475 (0.0827)	0.4959 (0.0749)	0.5483 (0.0794)	0.5021 (0.0785)	0.5790 (0.0801)
CS_9	0.1100 (0.0793)	0.2800 (0.0795)	0.4136 (0.0746)	0.4746 (0.0752)	0.4939 (0.0787)	0.5232 (0.0778)
CS_{10}	0.1875 (0.0733)	0.3156 (0.0779)	0.4378 (0.0740)	0.5255 (0.0768)	0.4966 (0.0795)	0.5623 (0.0804)
CS_{11}	0.1690 (0.0755)	0.4209 (0.0799)	0.5654 (0.0807)	0.6521 (0.0851)	0.6277 (0.0848)	0.7346 (0.0874)
CS_{12}	0.2078 (0.0736)	0.3232 (0.0789)	0.5424 (0.0777)	0.6369 (0.0789)	0.6322 (0.0834)	0.6600 (0.0836)
CS_{13}	0.0882 (0.0756)	0.3740 (0.0783)	0.4759 (0.0748)	0.5510 (0.0747)	0.5520 (0.0818)	0.6271 (0.0789)
CS_{14}	0.1722 (0.0739)	0.3304 (0.0778)	0.5082 (0.0805)	0.6344 (0.0806)	0.5544 (0.0846)	0.6768 (0.0851)
CS_{15}	0.1296 (0.0778)	0.3541 (0.0822)	0.5099 (0.0741)	0.5821 (0.0784)	0.5012 (0.0783)	0.6018 (0.0785)
CS_{16}	0.2083 (0.0724)	0.4173 (0.0805)	0.4877 (0.0727)	0.6310 (0.0743)	0.5844 (0.0763)	0.6790 (0.0779)
CS_{17}	0.3057 (0.0719)	0.4143 (0.0791)	0.5135 (0.0772)	0.6547 (0.0771)	0.6383 (0.0833)	0.6510 (0.0819)
CS_{18}	0.2220 (0.0726)	0.3725 (0.0833)	0.5763 (0.0740)	0.6595 (0.0773)	0.6629 (0.0811)	0.7068 (0.0818)
CS_{19}	0.2894 (0.0723)	0.4594 (0.0814)	0.5716 (0.0793)	0.6037 (0.0823)	0.6639 (0.0834)	0.6598 (0.0854)
CS_{20}	0.1190 (0.0738)	0.3012 (0.0755)	0.4155 (0.0720)	0.5325 (0.0744)	0.4978 (0.0752)	0.5597 (0.0751)
CS_{21}	0.3328 (0.0727)	0.4708 (0.0784)	0.6300 (0.0749)	0.6551 (0.0791)	0.6630 (0.0817)	0.7130 (0.0850)
CS_{22}	0.2435 (0.0732)	0.3575 (0.0794)	0.4839 (0.0692)	0.5662 (0.0735)	0.5327 (0.0754)	0.6373 (0.0777)
CS_{23}	0.2329 (0.0742)	0.3434 (0.0792)	0.4567 (0.0725)	0.4929 (0.0743)	0.4986 (0.0765)	0.5244 (0.0788)
CS_{24}	0.2043 (0.0718)	0.4283 (0.0794)	0.4852 (0.0717)	0.6850 (0.0734)	0.6242 (0.0797)	0.6817 (0.0792)
CS_{25}	0.2136 (0.0718)	0.3757 (0.0785)	0.4713 (0.0724)	0.5873 (0.0757)	0.5718 (0.0791)	0.6722 (0.0807)
CS_{26}	0.1691 (0.0733)	0.4021 (0.0810)	0.4914 (0.0782)	0.6029 (0.0789)	0.5643 (0.0841)	0.6556 (0.0855)
Log-likelihood	-29675.4509	-26299.3682	-17938.5982	-16823.9430	-17423.4417	-16782.5229
Parameters	33	35	41	69	43	71

Figure 3. Scale dynamics under different model specifications



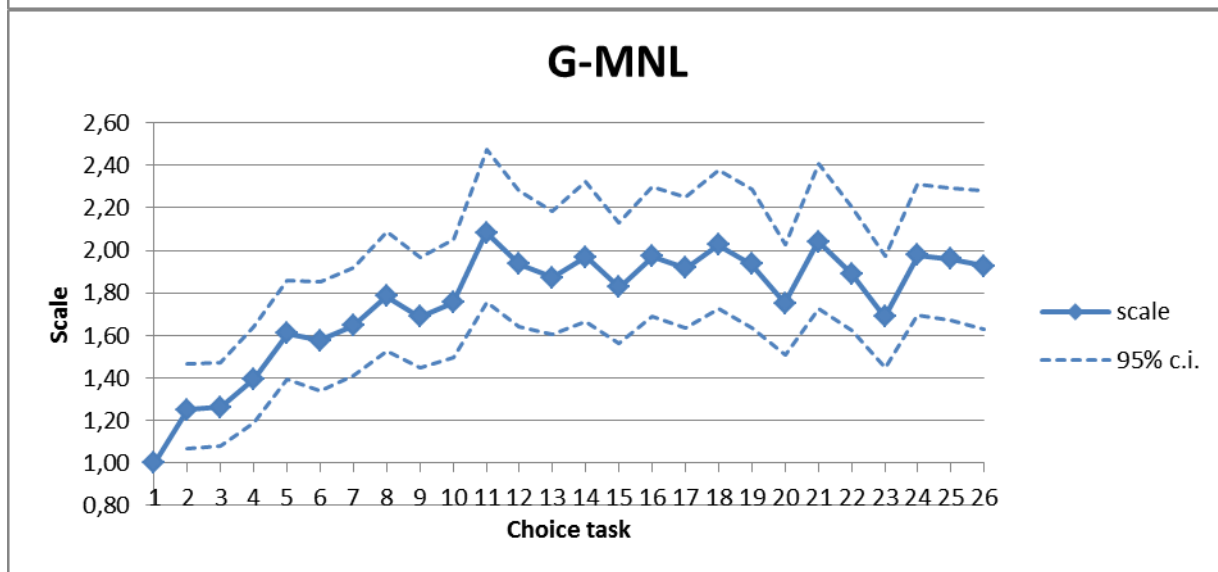
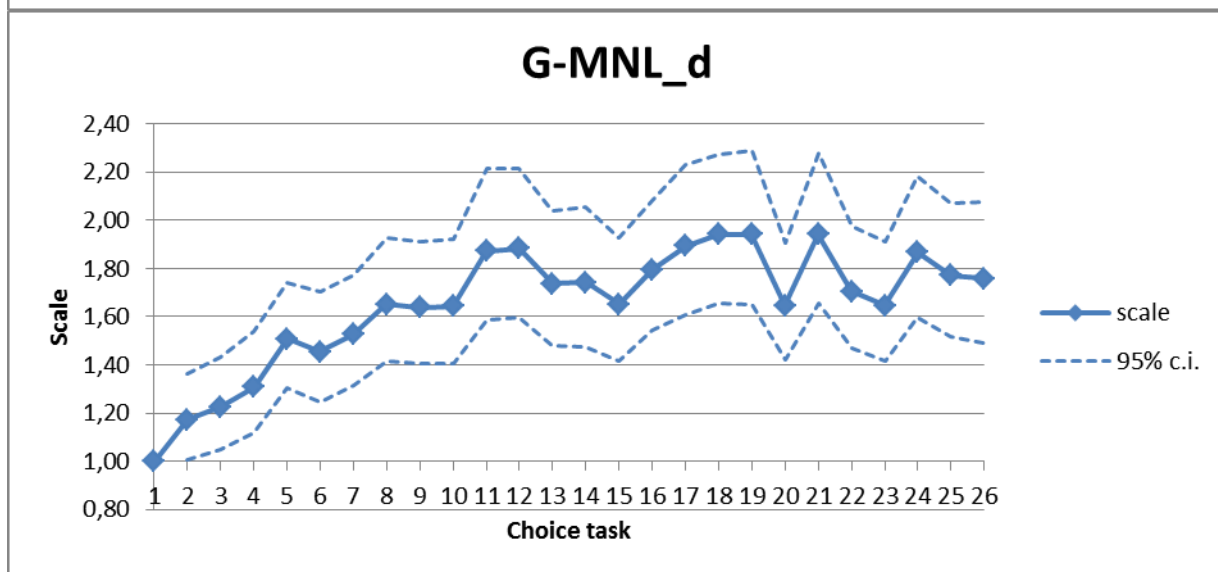
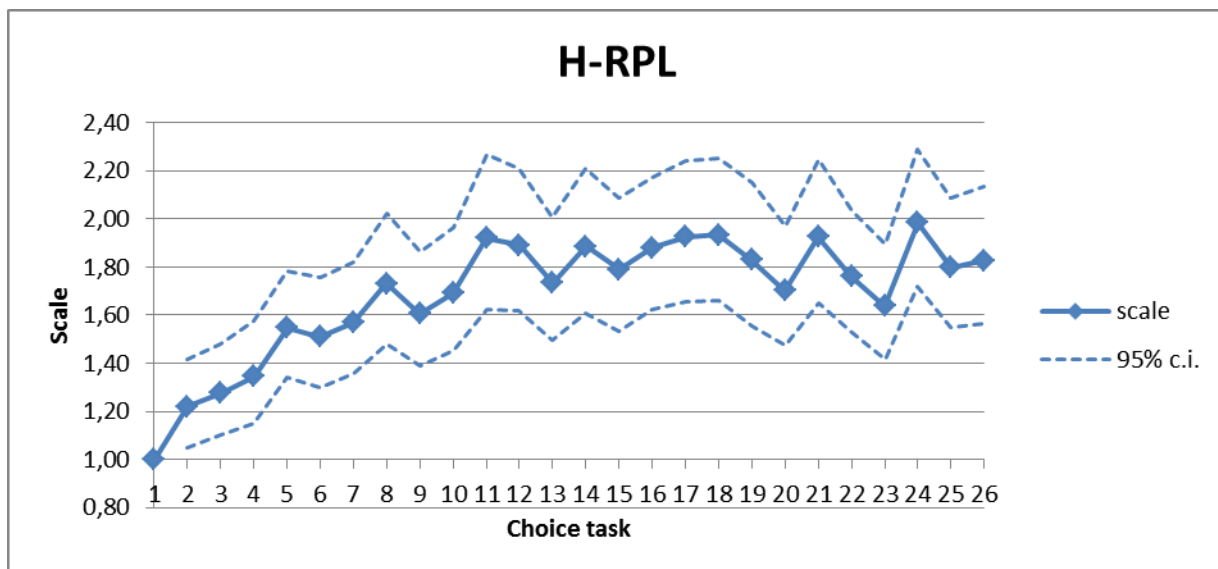
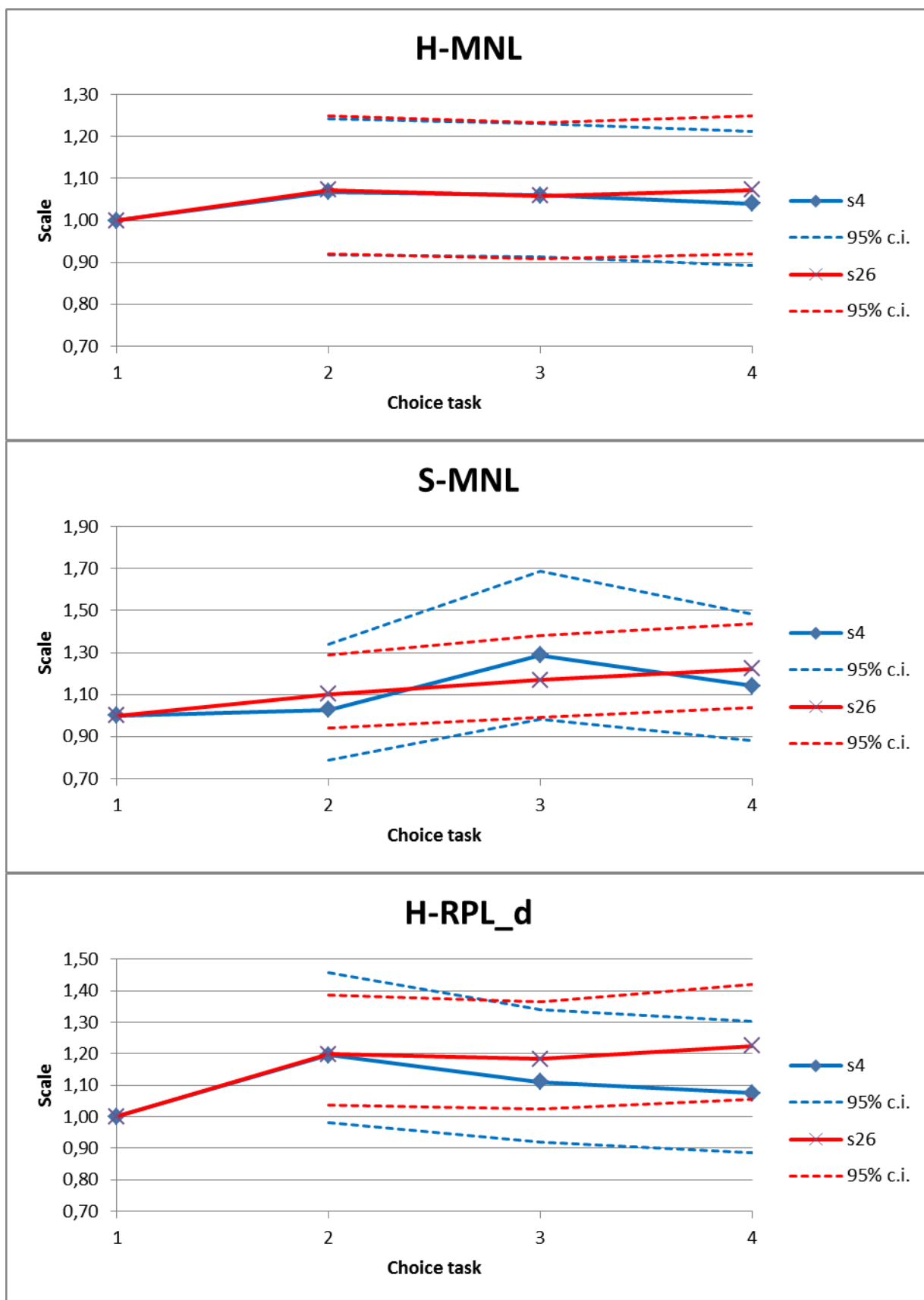
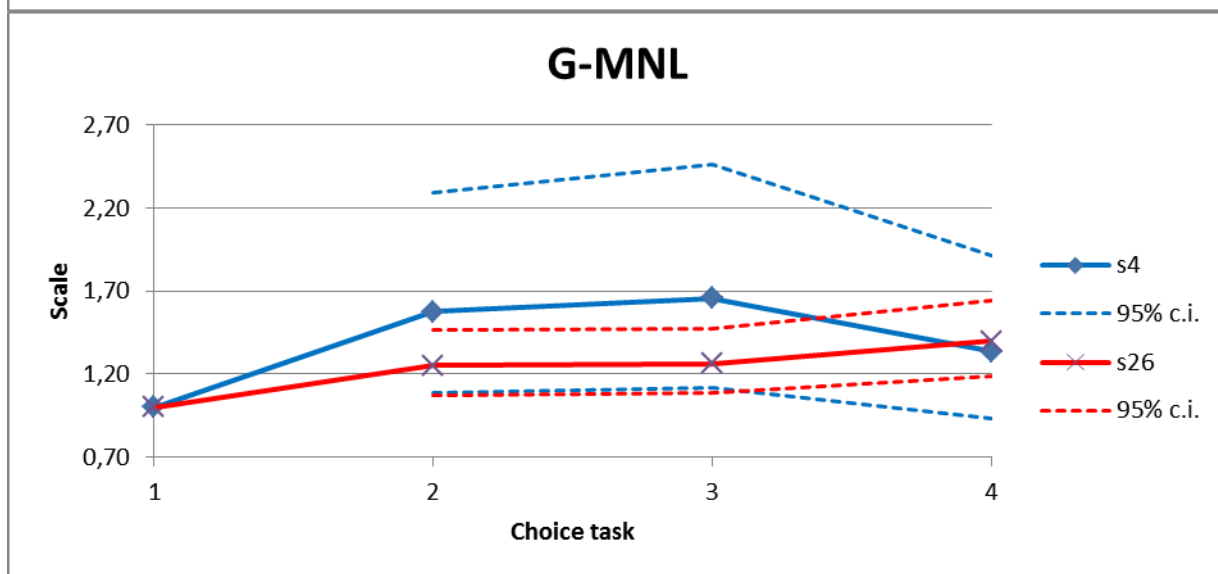
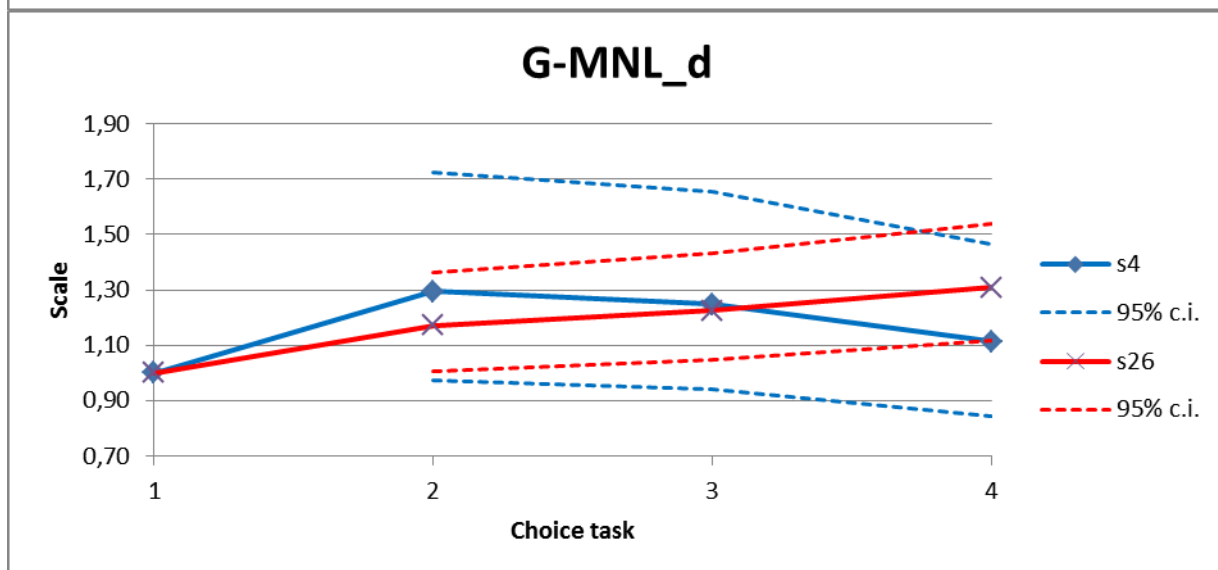
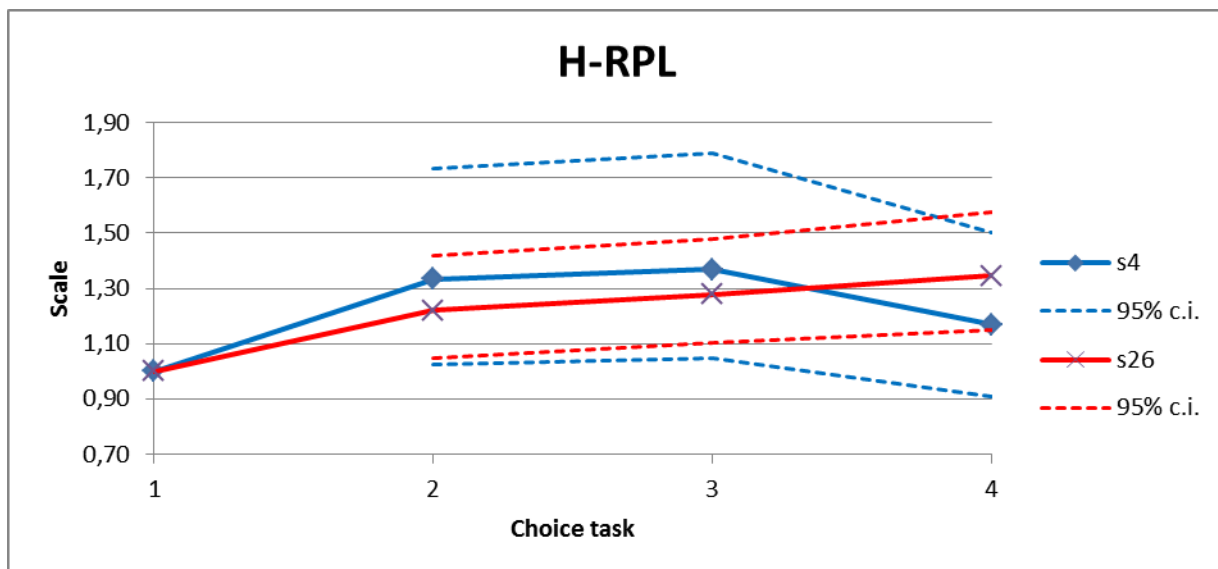


Figure 4. The influence of the number of choice tasks on confidence intervals of choice task specific scale parameters under different model specifications





5.3. Bias resulting from not accounting for ordering effects

Allowing for scale differences between choice tasks can substantially improve model fit, in addition to improvements resulting from allowing for unobservable preference and scale heterogeneity. We illustrate this with the results presented in Table 2. Every model used in our analysis was estimated with (panel 1) and without (panel 2) allowing for choice task specific scale. In the case of models allowing for preference or scale heterogeneity, allowing for scale dynamics by introducing 25 choice task specific dummy variables as covariates of scale significantly improves model performance, as represented by a statistically significant increase in model fit. This is not the case, however, for H-MNL model, where the increase in fit is not significant. This result is in line with our overall findings – not accounting for unobservable preference and scale heterogeneity may lead to underestimating scale dynamics and so controlling for them does not provide substantial improvement in the case too restrictive model is used (e.g. H-MNL).

Table 2. The influence of accounting for scale dynamics on model performance

	H-MNL	S-MNL	H-RPL_d	H-RPL	G-MNL_d	G-MNL
(1) With choice task specific scale						
No. of parameters	33	35	41	69	43	71
Log-likelihood	-29 675.45	-26 299.37	-17 938.60	-16 823.94	-17 423.44	-16 782.52
(2) Without choice task specific scale						
No. of parameters	8	10	17	44	18	46
Log-likelihood	-29 708.28	-26 349.26	-18 003.07	-16 949.04	-17 502.96	-16 853.54

Not accounting for differences in error term variance (scale) between choice tasks can result in biased estimates in a similar way as not accounting for panel structure of the dataset (Ortúzar & Willumsen 2001). One might think that since implicit prices are derived from the ratio of two parameters, scale cancels out and implicit prices are invariant to scale. This is not the case. Since the entire utility function (including scale) is used in estimation, and derived probabilities enter maximized log-

likelihood function, incorrect accounting for scale will influence parameter estimates, just as if the observations for different choice tasks were weighted in an arbitrary way.

In order to investigate the influence of this bias on implicit prices we used all the models discussed above to derive marginal WTP for changes in choice attributes. These results are presented in Table 3. The results show that even though there are some differences between implicit prices derived from different model specifications they are reasonably close. More importantly, it seems that not accounting for choice task specific scale does not lead to significant differences in implicit prices – they are well within confidence intervals. This result, however, may be specific to our dataset. Since we did not observe significant preference (taste) dynamics between choice tasks, not accounting for scale differences, and so effectively using an arbitrary weighting of data from different choice tasks, does not lead to changes in calculated implicit prices.

Table 3. Implicit prices under different model specifications [PLN]

	H-MNL	S-MNL	H-RPL_d	H-RPL	G-MNL_d	G-MNL
(1) With choice task specific scale						
NAT_1	59.66	54.28	30.65	31.57	32.44	29.03
NAT_2	87.75	74.63	45.09	42.33	46.22	39.10
TRA_1	107.25	69.43	42.97	37.47	42.28	34.41
TRA_2	143.20	104.74	62.71	55.75	62.61	49.98
INF_1	48.43	34.42	21.24	19.64	21.58	17.58
INF_2	78.24	50.80	31.13	27.47	30.68	24.85
(2) Without choice task specific scale						
NAT_1	59.32	53.88	32.54	33.79	30.78	32.96
NAT_2	87.28	74.24	46.84	46.27	44.70	44.62
TRA_1	106.68	69.15	45.51	40.79	41.57	38.28
TRA_2	142.71	104.53	65.82	60.27	60.66	56.53
INF_1	48.56	34.76	22.63	21.39	20.92	19.69
INF_2	78.24	50.87	32.77	30.66	30.86	27.68

6. Discussion and conclusions

In this study we investigated the dynamics of utility function parameters and its random component variance (scale), which may arise in repeated choice tasks of discrete choice experiments. Our study demonstrates that allowing for unobservable preference and scale heterogeneity substantially increases the magnitude of scale changes. We show how overly restrictive specifications of the model, especially when combined with not enough choice tasks to conduct meaningful analysis, may lead to drawing incorrect conclusions in terms of whether scale dynamics are observed. We argue that this is an important factor that allows to explain some of the contradicting evidence presented in existing empirical studies.

Our analysis utilized the state-of-the-art methods to control for unobservable preference and scale heterogeneity. We proposed extensions of these methods to allow the scale to be choice task specific. We find that the more sophisticated model is used, in terms of allowing for unobservable preference and scale heterogeneity, the higher observed increase in choice task specific scale parameter is. Using H-MNL model as a reference, we show that the magnitude of this increase grows from allowing for unobservable scale heterogeneity only (S-MNL) through allowing for unobservable preference heterogeneity (H-RPL) to allowing for both (G-MNL).

Empirically, we do not find evidence of statistically significant preference (taste) dynamics. However, we observe significant dynamics of scale – the scale appears to increase until about 8th choice task and then stabilizes. Therefore, as our respondents' choices became more deterministic with the number of completed choice tasks, we can conclude that we found evidence of learning. On the contrary, there are no signs of scale decrease, at least for the 26 choice tasks used in our study. This might be interpreted as no evidence of fatigue.

Our results do not support the inverted U-shaped relationship between scale and choice task number. In our case choice task specific scale was increasing at a decreasing rate, until after about 8th choice tasks it became relatively stable (with some natural variability). We note, however, that virtually all

studies that reported the inverted U-shaped relationship between scale and choice task number assumed a quadratic relationship rather than allowing scale to be choice task specific. As a result, their finding of scale reaching its maximum at 10 (Caussade *et al.* 2005) or 6 (Chung *et al.* 2010) choice tasks may be a result of assumed functional form of the relationship, combined with not enough choice tasks used in the studies.

Investigating the presence of fatigue is not straightforward, however. Swait and Adamowicz (2001b) note, that due to perceived high complexity of the first few choice-situations, or cumulative cognitive burden, respondents may simplify their choice strategies adopting noncompensatory decision rules. As a result one would observe increase (not decrease) in scale as choices would become more deterministic. Our dataset does not allow to verify this hypothesis. However, such a change in decision strategies would likely lead to the changes in estimated choice task specific preference (taste) parameters. Since in our case there was no evidence of such preference changes, we take it as an indication that such a change in decision strategies did not take place.

The economic theory of value maximization assumes that consumers have constant utility functions that are revealed when elicited (von Neumann & Morgenstern 1944). However, there are at least three streams of research which indicate that this might not be the case – both in terms of parameters of utility functions and in terms of its scale.

Following Simon's (1955) questioning of full rationality of human behavior, Heiner (1983) incorporated the notion of information processing limitations in the consumer's ability of making rational choices. These limitations could cause the respondents' decisions to change as they become more familiar with the 'institution' of DCE choice tasks (institutional learning; Braga & Starmer 2005). Respondents may also be expected to make errors, possibly increasingly as they become fatigued or bored in later choice tasks. Finally, if the choice tasks are complex, respondents may engage in simplification strategies, and change the choice strategies they use as fatigue or boredom sets in.

The second stream of research indicating problems with consumer' utility functions stability comes from the research of human decision making. There is some evidence that preferences may be

constructed, rather than revealed (Payne *et al.* 1992; Slovic 1995). This may mean that consumers could be learning about their true preferences throughout the course of a DCE study (*value learning*; Plott 1999) or that after a period of ‘burn-in’ choice-situations individuals evolve a systematic approach to evaluating alternatives (Luce & Tukey 1964), in either case resulting in changes in their underlying decision rules from one choice task to another. One other manifestation of constructed preferences would be that choices may be influenced by a number of ‘external’ stimuli, such as changes in the task environment (Payne *et al.* 1993). This is a yet another manifestation of the well-known *framing effect* (Kahneman & Tversky 2000). Finally, the constructed preference hypothesis may also cause path-dependence of respondent’s choices. For instance, constructed preferences may depend on the attribute levels seen in the opening or all previous choice tasks, possibly with decreasing strength. These may also be seen as a form of the *anchoring effect* (Kahneman *et al.* 1982). Even if individual’s choices are internally coherent, they may still be anchored to some reference point, such as the first choice task (*coherent arbitrariness*; Ariely *et al.* 2003).

Thirdly, in addition to behavioral reservations, there may be economic-theoretic reasons to why the choice outcomes may change with the progress of DCE study. Carson and Groves (2007) show that the conditions under which the single-bounded referendum format elicitation questions of contingent valuation studies are incentive compatible (take-it-or-leave it form of the question, respondents’ perception of consequentiality of their responses, possibility of introducing compelling payments at the stated price) may not hold under the DCE studies, in particular ones with repeated choice tasks. Since respondents may be aware in advance of having multiple choice tasks, and as they go they can exploit information about previous choice tasks and decisions, it remains unknown if the mechanism is incentive incompatible, and hence if there is the same decision mechanism underlying choices in all choice tasks. Evidence of such lag effects is provided by e.g. Holmes and Boyle (2005).

In the case of our study, using counterbalanced design allows to rule out systematic effect of many of the choice set specific effects discussed above, such as anchoring, framing or acting strategically. Since the order in which the 26 choice sets of our design were presented to different respondents was random, even if such effects as starting point bias or anchoring to a previous choice task attribute

levels were present, our approach causes these effects to cancel out between respondents. This allows us to focus on the effects of learning and fatigue in different choice tasks.

From the statistical point of view, it has long been recognized that treating repeated choice data in the same way as cross-sectional data may not be appropriate, and may lead to, among others, biased standard errors of the estimated parameters (Ortúzar & Willumsen 2001). Theoretically, the same holds if choice task specific scale differences are not accounted for. We investigated the magnitude of this bias on implicit prices under different model specifications. In the case of our study, even though accounting for choice task specific scale proved to significantly improve models fit, we did not observe any statistically significant differences in implicit prices if this was not accounted for. We note, however, that this effect can be specific to our study, as we did not observe significant changes in preference (taste) dynamics between choice tasks.

We acknowledge that our empirical results relate to a single dataset. We used data from a large-scale, high-quality national representative study conducted using face-to-face computer-assisted surveying techniques. It seems that, at least in the case of our study, the scale stabilized after about 8 choice tasks and remained on this level until the last, 26th choice task. This result can depend on many study-specific factors, however, such as context of the study, number of alternatives, number of attributes, and so on. There is also some evidence that other surveying techniques, such as internet-based surveys, might result in weaker learning effects and much stronger fatigue effects (Savage & Waldman 2008).

In conclusion, our study lays foundations for future research on ordering effects. Our analysis shows that investigating ordering effects requires assuring correct model specification, in particular allowing for unobserved preference and scale heterogeneity, as well as using enough choice tasks to provide meaningful conclusions. We illustrate how overly restrictive specification of the model, as well as using not enough choice tasks may lead to drawing incorrect conclusions in terms of whether scale dynamics are observed. This may prove an important factor in explaining some of the contradicting evidence presented in existing empirical studies. Finally, we show that accounting for choice task

specific scale dynamics can significantly improve model fit and urge other researchers to take it into account in their applications.

References

- Arentze T., Borgers A., Timmermans H. & DelMistro R. (2003). Transport stated choice responses: effects of task complexity, presentation format and literacy. *Transportation Research Part E: Logistics and Transportation Review*, 39, 229-244.
- Ariely D., Loewenstein G. & Prelec D. (2003). "Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences. *Quarterly Journal of Economics*, 118, 73-105.
- Bateman I.J., Burgess D., Hutchinson W.G. & Matthews D.I. (2008a). Learning design contingent valuation (LDCV): NOAA guidelines, preference learning and coherent arbitrariness. *Journal of Environmental Economics and Management*, 55, 127-141.
- Bateman I.J., Carson R.T., Day B., Dupont D., Louviere J.J., Morimoto S., Scarpa R. & Wang P. (2008b). Choice Set Awareness and Ordering Effects in Discrete Choice Experiments. In: CSERGE Working Paper EDM 08-01.
- Ben-Akiva M. & Morikawa T. (1990). Estimation of travel demand models from multiple data sources. In: *11'th International Symposium on Transportation and Traffic Theory Yokohama*.
- Bjornstad D., Cummings R. & Osborne L. (1997). A Learning Design for Reducing Hypothetical Bias in the Contingent Valuation Method. *Environmental and Resource Economics*, 10, 207-221.
- Bliemer M.C.J., Rose J.M. & Hess S. (2008). Approximation of Bayesian Efficiency in Experimental Choice Designs. *Journal of Choice Modelling*, 1, 98-127.
- Bradley M. & Daly A. (1992). Estimation of logit choice models using mixed stated preference and revealed preference information. In: *6'th International Conference on Travel Behaviour Quebec*.
- Bradley M. & Daly A. (1994). Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation*, 21, 167-184.
- Braga J. & Starmer C. (2005). Preference Anomalies, Preference Elicitation and the Discovered Preference Hypothesis. *Environmental and Resource Economics*, 32, 55-89.
- Brazell J. & Louviere J. (1997). Respondent's Help, Learning and Fatigue. In: *INFORMS Marketing Science Conference University of California, Berkeley*.
- Brouwer R., Dekker T., Rolfe J. & Windle J. (2010). Choice Certainty and Consistency in Repeated Choice Experiments. *Environmental and Resource Economics*, 46, 93-109.
- Carlsson F. & Martinsson P. (2001). Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments?: Application to the Valuation of the Environment. *Journal of Environmental Economics and Management*, 41, 179-192.
- Carlsson F., Raun Mørkbak M. & Bøye Olsen S. (2010). The first time is the hardest: A test of ordering effects in choice experiments. In: Working Papers in Economics no. 470.
- Carson R. & Groves T. (2007). Incentive and informational properties of preference questions. *Environmental and Resource Economics*, 37, 181-210.
- Caussade S., Ortúzar J.d.D., Rizzi L.I. & Hensher D.A. (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation Research Part B: Methodological*, 39, 621-640.
- Chung C., Boyer T. & Han S. (2010). How many choice sets and alternatives are optimal? Consistency in choice experiments. *Agribusiness*, 27, 114-125.
- Day B., Bateman I.J., Carson R.T., Dupont D., Louviere J.J., Morimoto S., Scarpa R. & Wang P. (2012). Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of Environmental Economics and Management*, 63, 73-91.
- Day B. & Pinto P.J.-L. (2010). Ordering anomalies in choice experiments. *Journal of Environmental Economics and Management*, 59, 271-285.
- Dellaert B.G.C., Brazell J.D. & Louviere J.J. (1999). The Effect of Attribute Variation on Consumer Choice Consistency. *Marketing Letters*, 10, 139-147.
- DeSarbo W.S., Lehmann D.R. & Hollman F.G. (2004). Modeling Dynamic Effects in Repeated-Measures Experiments Involving Preference/Choice: An Illustration Involving Stated Preference Analysis. *Applied Psychological Measurement*, 28, 186-209.

- Ferrini S. & Scarpa R. (2007). Designs with a priori information for nonmarket valuation with choice experiments: A Monte Carlo study. *Journal of Environmental Economics and Management*, 53, 342-363.
- Fiebig D.G., Keane M.P., Louviere J. & Wasi N. (2010). The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity. *Marketing Science*, 29, 393-421.
- Fosgerau M. (2006). Investigating the distribution of the value of travel time savings. *Transportation Research Part B: Methodological*, 40, 688-707.
- Fosgerau M. & Nielsen S.F. (2010). Deconvoluting Preferences and Errors: A Model for Binomial Panel Data. *Econometric Theory*, 26, 1846-1854.
- Haaaijer R., Kamakura W. & Wedel M. (2000). Response Latencies in the Analysis of Conjoint Choice Experiments. *Journal of Marketing Research*, 37, 376-382.
- Hanley N., Wright R.E. & Koop G. (2002). Modelling Recreation Demand Using Choice Experiments: Climbing in Scotland. *Environmental and Resource Economics*, 22, 449-466.
- Heiner R.A. (1983). The Origin of Predictable Behavior. *The American Economic Review*, 73, 560-595.
- Hensher D. & Greene W. (2003). The Mixed Logit model: The state of practice. *Transportation*, 30, 133-176.
- Hensher D., Louviere J. & Swait J. (1998). Combining sources of preference data. *Journal of Econometrics*, 89, 197-221.
- Hensher D.A. (2001). Measurement of the Valuation of Travel Time Savings. *Journal of Transport Economics and Policy*, 35, 71-98.
- Hensher D.A. (2004). Identifying the Influence of Stated Choice Design Dimensionality on Willingness to Pay for Travel Time Savings. *Journal of Transport Economics and Policy*, 38, 425-446.
- Hensher D.A. (2006a). How do respondents process stated choice experiments? Attribute consideration under varying information load. *Journal of Applied Econometrics*, 21, 861-878.
- Hensher D.A. (2006b). Revealing Differences in Willingness to Pay due to the Dimensionality of Stated Choice Designs: An Initial Assessment. *Environmental and Resource Economics*, 34, 7-44.
- Hensher D.A. & Rose J.M. (2005). Respondent Behavior in Discrete Choice Modeling with a Focus on the Valuation of Travel Time Savings. *Journal of Transportation and Statistics*, 8, 17-30.
- Hess S., Hensher D.A. & Daly A. (2012). Not bored yet – Revisiting respondent fatigue in stated choice experiments. *Transportation Research Part A: Policy and Practice*, 46, 626-644.
- Hess S., Rose J.M. & Polak J. (2010). Non-trading, lexicographic and inconsistent behaviour in stated choice data. *Transportation Research Part D: Transport and Environment*, 15, 405-417.
- Hess S. & Train K. (2011). Recovery of inter- and intra-personal heterogeneity using mixed logit models. *Transportation Research Part B: Methodological*, 45, 973-990.
- Holmes T.P. & Boyle K.J. (2005). Dynamic Learning and Context-Dependence in Sequential, Attribute-Based, Stated-Preference Valuation Questions. *Land Economics*, 81, 114-126.
- Kahneman D., Slovic P. & Tversky A. (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.
- Kahneman D. & Tversky A. (2000). *Choices, Values, and Frames*. Cambridge University Press, Cambridge, UK.
- Keppel G. & Wickens T.D. (2004). *Design and Analysis: A Researcher's Handbook*. Prentice Hall.
- Ladenburg J. & Olsen S.B. (2008). Gender-specific starting point bias in choice experiments: Evidence from an empirical study. *Journal of Environmental Economics and Management*, 56, 275-285.
- Liechty J.C., Fong D.K.H. & DeSarbo W.S. (2005). Dynamic Models Incorporating Individual Heterogeneity: Utility Evolution in Conjoint Analysis. *Marketing Science*, 24, 285-293.
- Louviere J., Street D., Carson R., Ainslie A., Deshazo J.R., Cameron T., Hensher D., Kohn R. & Marley T. (2002). Dissecting the Random Component of Utility. *Marketing Letters*, 13, 177-193.
- Luce R.D. & Tukey J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1-27.
- McFadden D. (1986). The Choice Theory Approach to Market Research. *Marketing Science*, 5, 275-97.

- McFadden D. & Train K. (2000). Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, 15, 447-470.
- Oppewal H., Morrison M., Wang P. & Waller D. (2010). Preference Stability: Modeling how COConsumer Preferences Shift after Receiving New Product Information. In: *Choice Modelling: The State-of-the-art and the State-of-practice* (eds. Hess S & Daly A). Emerald Group Publishing Limited.
- 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. In: *Paper Presented at the European Transport Conference* Glasgow.
- Ortúzar J.d.D. & Willumsen L.G. (2001). *Modelling Transport*. Wiley.
- Palma A.d., Myers G.M. & Papageorgiou Y.Y. (1994). Rational Choice Under an Imperfect Ability To Choose. *The American Economic Review*, 84, 419-440.
- Payne J.W., Bettman J.R. & Johnson E.J. (1992). Behavioral Decision Research: A Constructive Processing Perspective. *Annual Review of Psychology*, 43, 87-131.
- Payne J.W., Bettman J.R. & Johnson E.J. (1993). *The Adaptive Decision Maker*. Cambridge University Press.
- Plott C.R. (1999). Rational Individual Behavior in Markets and Social Choice Processes. In: *The Rational Foundations of Economic Behavior* (eds. Arrow KJ, Colombatto E, Perlman M & Schmidt C). Palgrave Macmillan, pp. 225-250.
- Revelt D. & Train K. (1998). Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level. *Review of Economics and Statistics*, 80, 647-657.
- Rose J. & Black I. (2006). Means matter, but variance matter too: Decomposing response latency influences on variance heterogeneity in stated preference experiments. *Marketing Letters*, 17, 295-310.
- Rose J.M., Hensher D.A., Caussade S., Ortúzar J.d.D. & Jou R.-C. (2009). Identifying differences in willingness to pay due to dimensionality in stated choice experiments: a cross country analysis. *Journal of Transport Geography*, 17, 21-29.
- Ruud P.A. (1996). Approximation and Simulation of the Multinomial Probit Model: An Analysis of Covariance Matrix Estimation. In. working paper, Department of Economics, University of California, Berkeley.
- Sándor Z. & Wedel M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, 38, 430-444.
- Savage S.J. & Waldman D.M. (2008). Learning and fatigue during choice experiments: a comparison of online and mail survey modes. *Journal of Applied Econometrics*, 23, 351-371.
- Scarpa R. & Rose J.M. (2008). Design Efficiency for Non-Market Valuation with Choice Modelling: How to Measure it, What to Report and Why. *Australian Journal of Agricultural and Resource Economics*, 52, 253-282.
- Scheufele G. & Bennett J. (2010). Effects of alternative elicitation formats in discrete choice experiments. In: *Australian Agricultural and Resource Economics Society conference* Adelaide, Australia.
- Simon H.A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69, 99-118.
- Slovic P. (1995). The Construction of Preference. *American Psychologist*, 50, 364-371.
- Street A.P. & Street D.J. (1987). *Combinatorics of Experimental Design*. Oxford University Press.
- Swait J. & Adamowicz W. (2001a). Choice Environment, Market Complexity, and Consumer Behavior: A Theoretical and Empirical Approach for Incorporating Decision Complexity into Models of Consumer Choice. *Organizational Behavior and Human Decision Processes*, 86, 141-167.
- Swait J. & Adamowicz W. (2001b). The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching. *The Journal of Consumer Research*, 28, 135-148.
- Swait J. & Louviere J. (1993). The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. *Journal of Marketing Research*, 30, 305-314.
- von Neumann J. & Morgenstern O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.

Annex 1. Review of empirical studies investigating ordering effects

Study	Attributes	Alternatives	Choice tasks	Sample size	Rotation	Administration	Context	Methodology	Unobservable preference heterogeneity	Unobservable scale heterogeneity	Findings – preference changes	Findings – scale changes
Bjornstad <i>et al.</i> (1997)	–	–	1,3 x CVM	609	no	lab	nature conservation	Non-parametric	no	no	learning, reduction in hypothetical bias	–
Hensher (2001)	6	2	16	198	yes	CAPI	transport (travel choice)	Compared WTP for 1-4, 8, 12, 16 CT and 1-4, 5-8, 9-12, 13-16 CT	yes	some	no effect	–
Swait and Adamowicz (2001a)	5	4	16	280	no	n/a	food choice	Latent class model with CT number (and cumulative complexity) as class membership variables	no	no	Respondents simplify strategies – adopt simpler decision rules (especially after CT 8)	–
Hanley <i>et al.</i> (2002)	6	3	4, 8	367	no	mail	climbing route	Separate multinomial logit and nested logit models for 4 and 8 CTs	no	no	no effect	–
DeSarbo <i>et al.</i> (2004)	10	–	27+3	162	yes	students	students' apartments	ratings dependent on order	no	no	Adaptation / evolution of utility function as respondents progress through CTs	–
Bateman <i>et al.</i> (2008a)	n/a	n/a	1,4 x CVM	400	no	FTF	animal welfare	DB, compare WTP based on 1 and 2 bid answer	no	no	experience reduces SB/DB differences (institutional learning) and reduces anchoring to the first bid (value learning, no coherent arbitrariness)	–
Ladenburg and Olsen (2008)	5	3	6	294+285	no	CAWI	nature protection	Swait-Louviere procedure for 3+3CTs	no	no	Some learning (gender specific); starting point bias (decaying effect)	–

Rose <i>et al.</i> (2009)	3-6	3-5	6-15	501	no	CAPI	transport (route choice)	implicit prices a function of no of choice tasks in the design	yes	no	Evidence mixed, country specific	–
Caussade <i>et al.</i> (2005)				259								
Hensher (2004, 2006b, a)				427								
Bradley and Daly (1994)	4	2	10-16	243	yes	CAPI	transport (route choice)	Logit scaling approach	no	no	–	fatigue
Brazell and Louviere (1997)	10	3	12-96	553	yes	mail	holiday activities	Swait-Louviere procedure, between and within surveys with different number of CTs	no	no	–	No effect
	7	3	16-120	224	n/a	mail	canned soup choice					
Carlsson and Martinsson (2001)	3	2	16+16	35	No	lab	donations to WWF	Swait-Louviere procedure	no	no	Mixed, Non-constant	n/a
Arentze <i>et al.</i> (2003)	3,5	2,3	8+8	344	yes	n/a	transport (route choice)	Logit scaling approach	no	no	No effect	No effect
Caussade <i>et al.</i> (2005)	3-6	3-5	6-15	403	no	CAPI	transport (route choice)	Heteroskedastic (covariance heterogeneity)logit, scale modeled as a function of design dimensions	no	no	–	U-shaped relationship, max scale for 10 th CT
Holmes and Boyle (2005)	7	2	4	926	no	mail	forest management	Swait-Louviere procedure	no	some	Structural change between 1-3 and 4 CT	n/a
Savage and Waldman (2008)	5	2	8	357+325	yes	Mail + CAWI	internet provider service	Error components model	no	yes	–	mail – no effect CAWI – fatigue
Bateman <i>et al.</i> (2008b)	3	2	16	864	yes	FTF	drinking water quality	1 st CT repeated at the end – compare choices and WTP; include interaction of price with log-order	yes	some	Implicit prices decreasing with CT number; differences between the 1 st and the 16 th (repeated 1 st) CT not statistically significant	Effect not statistically significant
Oppewal <i>et al.</i> (2010)	8	3	4+4+4	400	no	CAWI	marketing (DVD recorders)	Heteroscedastic logit model, compare scale between CT 5-8 and CT 9-12	no	no	n/a	No effect
Scheufele and Bennett (2010)	3	2	1,4	1444+367+371+369+376	no	CAWI	nature conservation	Swait-Louviere procedure, multinomial logit model	no	no	1 CT questionnaire yields significantly higher implicit prices than 4 CT	Learning
Brouwer <i>et al.</i> (2010)	3	3	6	300	yes	FTF	water scarcity	Swait-Louviere, random parameters logit model	yes	some	No effect	No effect

Carlsson <i>et al.</i> (2010)	5	2	8+8	389	no	CAWI	food choice	Swait-Louviere procedure	yes	no	Mixed	No effect if CT 2-9 vs. 9-16 compared, when CT 1 included – learning
Chung <i>et al.</i> (2010)	7	3-12	1-20	1000	yes	FTF	food choice	Heteroskedastic (covariance heterogeneity) logit, scale modeled as a function of design dimensions	no	no	–	U-shaped relationship, max scale for 6'th CT
Hess <i>et al.</i> (2012)	2	3	8	1563+1146+1110	Yes	CAWI	transport (route choice)	Logit scaling approach, MNL and RPL model	yes	no	mixed	mixed, some learning, no evidence of fatigue
Orr <i>et al.</i> (2010)	2	2	5+5+5	397	Yes	CAPI	transport safety					
Fosgerau (2006)	2	2	8	472+1725	Yes	CAPI	transport (route choice)					
Hensher and Rose (2005)	5	3	16	237+205	Yes	CAPI	transport (route choice)					
Hess <i>et al.</i> (2010)	5	3	16	304	yes	CAPI	transport (route choice)					
Day <i>et al.</i> (2012)	3	2	17	864	yes	FTF	tap water quality	Non parametric + random effects probit with random scale	some	yes	precedent-dependent order effects	no effect

Annex 2a – the illustration of economic forest and close-to-natural forest



Economic forest



Close-to-natural forest

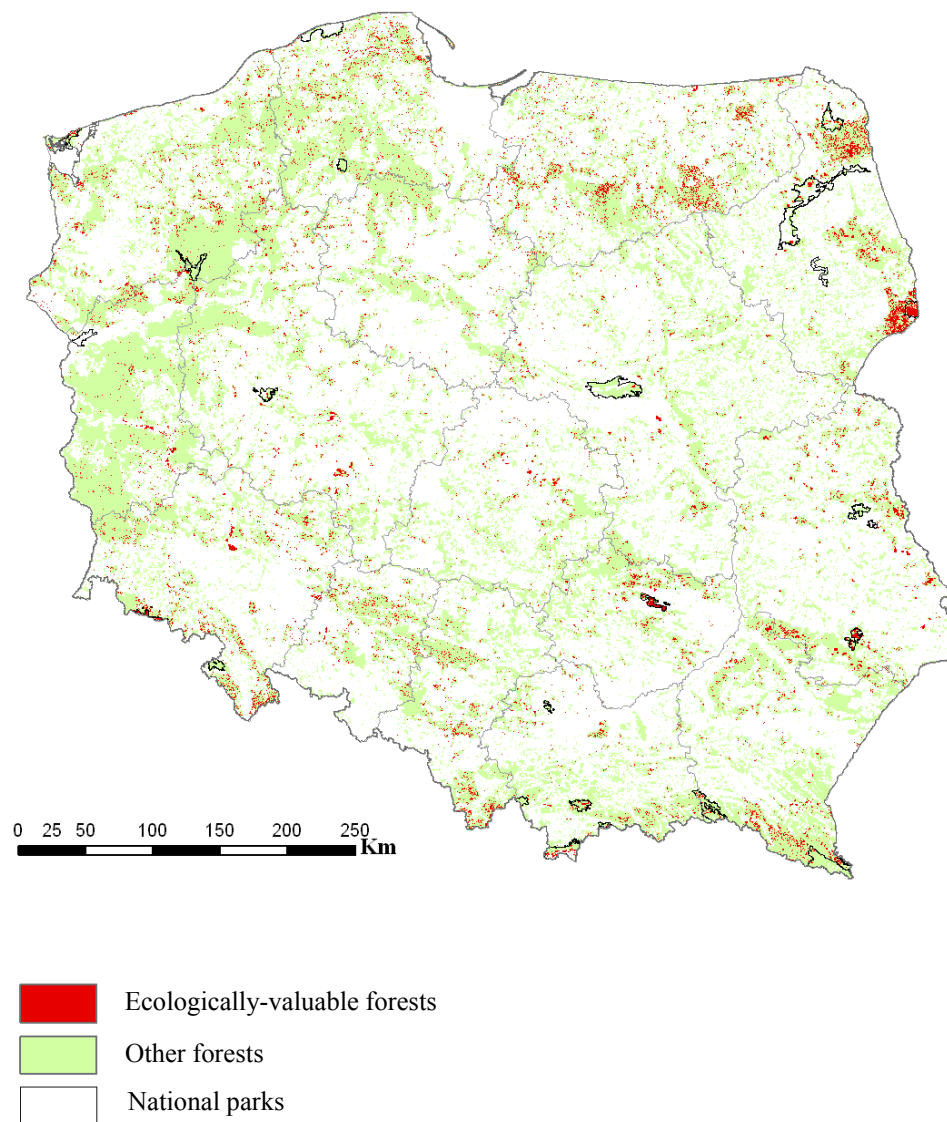
Annex 2b – the illustration of litter in the forests



Annex 2c – the illustration of tourist infrastructure



Annex 3 – The most ecologically valuable forests in Poland





FACULTY OF ECONOMIC SCIENCES
UNIVERSITY OF WARSAW
44/50 DŁUGA ST.
00-241 WARSAW
WWW.WNE.UW.EDU.PL