WORKING PAPER

ITLS-WP-13-13

Estimation of stochastic scale with best-worst data.

By
Andrew T. Collins and John M. Rose

July 2013

ISSN 1832-570X

INSTITUTE of TRANSPORT and LOGISTICS STUDIES

The Australian Key Centre in
Transport and Logistics Management

The University of Sydney

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

NUMBER: Working Paper ITLS-WP-13-13

TITLE: **Estimation of stochastic scale with best-worst data.**

ABSTRACT: Recently there has been a steady stream of literature advocating the best-worst response mechanism, where respondents are asked to sequentially choose the best and worst alternatives in a choice set, resulting in a partial or complete ranking of the alternatives. In this paper, we present an empirical study in which respondents were encouraged to respond using repeated best-worst, but were nonetheless free to respond in any order they preferred. Models are estimated that account for three alternative response processes: conventional ranking of the alternatives from best to worst, sequential best-worst choice, and two best choices followed by two worst choices. While the sequential best-worst models perform best, the sensitivities retrieved are consistent across all three models. We find strong evidence of stochastic scale heterogeneity across respondents, where the extent of this heterogeneity is also consistent across all three model forms. However, deterministic scale heterogeneity, that accounts for differences in scale across each of the pseudo-observations, is not consistent across the model forms, with respect to the implied rank of the observation. Rather, the consistency is with the number of alternatives associated with the pseudo-observation, with scale decreasing as the number of alternatives decreases. A test of alternative specifications of the panel in the mixture model used to identify stochastic scale identifies that scale should be invariant across the full set of responses by an individual, rather than just the responses from each rank from that individual. Despite an overall finding that the sensitivities retrieved are robust to the assumption of the completion order of the ranking within the model, differences in sensitivities retrieved from each best-worst choice raise concerns with pooling the data across best-worst choices, in line with concerns raised previously with rankings data.

KEY WORDS: ***Best worst, scale, rank explosion, SMNL.***

AUTHORS: **Collins and Rose**

Acknowledgements: The authors would like the thank Ric Scarpa, who was involved in the design of the empirical study.

CONTACT:

INSTITUTE of TRANSPORT and LOGISTICS STUDIES (C13)
The Australian Key Centre in Transport and Logistics Management

The University of Sydney NSW 2006 Australia

Telephone: +612 9114 1824
Facsimile: +612 9114 1722
E-mail: business.itlsinfo@sydney.edu.au
Internet: http://sydney.edu.au/business/itls

DATE: July 2013

1. Introduction

Traditionally, analysts using stated choice type experiments to collect preference data have adopted the 'pick one' choice response approach in which respondents are asked to select their most preferred alternative out of the set presented to them (see Figure 1a for an example). Use of the pick one response mechanism is most commonly used due to the perception that such survey responses reflect real market outcomes, where decision makers are observed to make choices rather than rate or rank the available alternatives, and hence reflect a more natural and realistic type of response from the perspective of the respondent. From the analysts perspective however, the pick one response provides only limited information as to the underlying preference structure for the alternatives shown, as no information is captured on the relative desirability of the remaining non chosen alternatives. This limitation has led to a number of alternative response mechanisms being developed and tested within the literature. One such mechanism is to ask the respondent to provide a complete or partial ranking of the presented alternatives (Chapman and Staelin 1982), thus obtaining more information from each choice task (see Figure 2 for an example of a ratings task), however as stated above, this comes at the cost of being a less realistic task for the respondent. If a logit model is to be estimated, each ranking can be treated as an independent choice (or pseudo-observation) from all alternatives that remain after alternatives chosen for higher ranks are omitted (Luce and Suppes 1965), in an approach commonly referred to as exploded logit.

If you were looking through a dating website and had a choice among the three people shown based on the descriptions listed, which person would you choose to contact?

	Person A	Person B	Person C	None
Drinking Habit	Casual drinker	Non drinker	Moderate drinker	
Smoking Habit	Ex smoker	Smoker	Non smoker	
Children	Doesn't want children	Single parent	None currently	
Job	White Collar	Unemployed	Blue Collar	
Looks	Below average	Above average	Average	
Cost to contact	$20	$20	$10	
I would choose to contact	○	○	○	○

Next

Fig 1: 'Pick one' experiment

An alternative response mechanism that is becoming more popular amongst practitioners, particularly in the field of marketing, is the best-worst choice approach where the respondent is not asked to pick their single most preferred alternative or to rank or rate all alternatives shown, but rather to indicate the best and worst alternatives from the set shown (see Figure 3). In cases involving more than three alternatives present, additional questions about which of the remaining alternatives are considered next best and next worst may produce a full preference ranking (see Figure 4). Without these additional responses, partial preference rankings are obtained. Aside from providing additional preference information, the motivation for this approach over rankings or rating formats is that some evidence suggests that choosing extremes is cognitively less burdomsome (Louviere et al. 2008).

If you were looking through a dating website and considered contacting among the five people shown, based on the description, please rank from 1 to 5 your most preferred contact candidate, where 1 represents the person you would most prefer to contact and 5 the person you would least like to contact.

	Person A	Person B	Person C	Person D	Person E
Drinking Habit	Casual drinker	Non drinker	Moderate drinker	Non drinker	Moderate drinker
Smoking Habit	Non smoker	Smoker	Ex smoker	Non smoker	Smoker
Children	Doesn't want children	Doesn't want children	None currently	Single parent	None currently
Job	White Collar	Blue Collar	Unemployed	Unemployed	White Collar
Looks	Average	Below average	Above average	Above average	Below average
Cost to contact	$20	$15	$10	$10	$20
Ranking (1-5)					

Next

Fig 2: Ranking experiment

If you were looking through a dating website and considered contacting among the five people shown based on the descriptions listed, which profile represents the best candidate and which represents the worst? And then which is the best and which is the worst of the three remaining profiles?

	Person A	Person B	Person C	Person D	Person E
Drinking Habit	Non drinker	Moderate drinker	Non drinker	Casual drinker	Casual drinker
Smoking Habit	Ex smoker	Ex smoker	Smoker	Non smoker	Smoker
Children	None currently	None currently	Single parent	Single parent	None currently
Job	Unemployed	Blue Collar	Unemployed	White Collar	White Collar
Looks	Above average	Below average	Below average	Average	Average
Cost to contact	$15	$15	$15	$20	$15
Which profile do you consider to be the best and which is the worst?					

Next

Fig 3: Partial ranking best worst response format

Depending on how the survey is administered, respondents may be either constrained to answer following a structured pattern of choices such as what is the best alternative, what is the worst, what is the next best alternative, what is the next worst, etc. (e.g., Scarpa et al. 2009; Marley and Pihlens 2010), or be free to respond to the choice task in any order desired (e.g., one respondent may answer what is the best alternative, what is the next best, etc. whilst a second respondent may indicate what is the best alternative, followed by what is the worst, etc.). The data in the current study allows for the latter type of answers. Depending on how the questions have been answered, different modelling approaches may be advisable given that the process of answering the questions may reveal different preference structures, which may have implications as to how the data is treated for purposes of estimation. In the case of full rankings data being captured using the best-worst response format, it is possible to rank explode the data to obtain additional pseudo-choice-observations which can be used for model estimation. For example, assuming five alternatives A, B, C, D and E have been ranked from best to worst in the same order shown, the first choice observation may consist of all alternatives A, B, C, D and E. New pseudo-observations may then be constructed by eliminating the previously most preferred option, such that pseudo-observation 1 will consist of alternatives B, C, D and E, pseudo-observation 2 of C, D and E and pseudo-observation 3 of D and E. Alternatively, a formulation of the choice probability of choosing the worst alternative can be derived, with the choices then modelled as a succession of choices of the best and worst alternatives from the remaining alternatives (e.g., Lancsar and Louviere 2008;

Marley and Pihlens 2010). Marley and Pihlens (2010) suggest that the decision processes that the decision makers employ may impact on which models best fit the data, and in turn upon the modelled results.

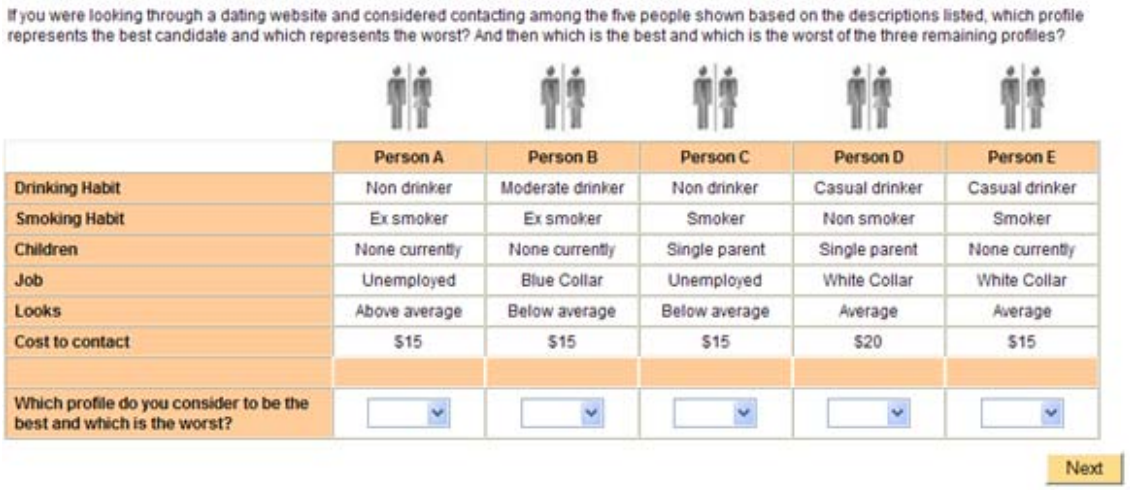

Fig 4: Full ranking best worst response format

One area shown to be particularly affected by how rankings data is treated is scale. Evidence from numerous studies involving rankings data suggests that the scale of the systematic component of utility (or equivalently, the magnitude of the unexplained variance) is not consistent across each of the pseudo-observations (i.e., the observations for each ranking). In the exploded logit framework, Hausman and Ruud (1987), Ben-Akiva et al. (1991) and Bradley and Daly (1994) demonstrated decreasing scale over increasing ranks, although the estimated decreases were not necessarily found to be monotonic. Such a finding could have a number of explanations, including a lack of engagement with the choice of lower ranks, and greater difficulty in choosing the lower ranks. Scarpa et al. (2009) also estimated an exploded logit model, with the explosion of choices assuming sequential best choices, but with sequential elicitation of best then worst choices in the study, over five alternatives. They identified higher scale for the first round of best-worst choices than for the second round, which suggests that the overall best and worst choices might be easier to make; thus the finding is supportive of the best-worst framework.

This paper presents three core contributions. Both between respondent and between rank scale differences are handled simultaneously, in a mixture model; the panel specification of this model is explored; and various models that handle different choice sequences are compared, notably conventional ranking, sequential best-worst, and all best choices followed by all worst choices. As such, we estimate and compare models estimated under three different assumptions of how the data structure should be analysed; a conventional exploded logit model, a model of repeated best then worst choices, and a model of two best choices followed by two worst choices. In all cases, we utilise the scaled multinomial logit (SMNL) model (Fiebig et al. 2010) which allows for a stochastic treatment of scale. The SMNL model also allows scale to be decomposed by other observed characteristics in the data such as socio-demographic or contextual effects. In terms of the panel nature of the data, in the context of exploded logit and best-worst data, no *a priori* hypothesis currently exists as to whether scale should be invariant across the full set of responses by an individual, or alternatively across just the responses from each rank from that individual. As such, we test both assumptions empirically herein.

The remainder of the paper is structured as follows. The next section details the methodology that will be employed, followed by an overview of the empirical setting. Next, results from the empirical study are presented after which discussion and conclusions drawn from the current study are presented.

2. Methodology

Under the conventional exploded logit approach, the probability of a particular ranking of alternatives can be expressed as the product of logit formulas. The assignment of each alternative to a rank by the respondent constitutes a choice of the best alternative from those that remain. Consider the probability P of a complete ranking across five alternatives A, ..., E, where, for example, the ranking of the alternatives from best to worst is A, B, C, D, E. The probability is given in Equation (1).

$$P(A,B,C,D,E) = \frac{e^{\beta x_A}}{\sum_{j=A,B,C,D,E} e^{\beta x_j}} \frac{e^{\beta x_B}}{\sum_{j=B,C,D,E} e^{\beta x_j}} \frac{e^{\beta x_C}}{\sum_{j=C,D,E} e^{\beta x_j}} \frac{e^{\beta x_D}}{\sum_{j=D,E} e^{\beta x_j}} . \quad (1)$$

Each of the four choices of rank can be considered as an independent 'pseudo-observation'. Equation (1) suggests that the scale of each pseudo-observation is identical, and normalised to one. This need not be the case, where for example a scale term μ_r could be parameterised and introduced for each rank *r*:

$$P(A,B,C,D,E) = \frac{e^{\mu_1 \beta x_A}}{\sum_{j=A,B,C,D,E} e^{\mu_1 \beta x_j}} \frac{e^{\mu_2 \beta x_B}}{\sum_{j=B,C,D,E} e^{\mu_2 \beta x_j}} \frac{e^{\mu_3 \beta x_C}}{\sum_{j=C,D,E} e^{\mu_3 \beta x_j}} \frac{e^{\mu_4 \beta x_D}}{\sum_{j=D,E} e^{\mu_4 \beta x_j}} . \quad (2)$$

So long as one of μ_r is normalised (setting $\mu_1 = 1$ is perhaps most intuitive), the model is identified, and the possibility exists for differences in scale across the ranking choices to be observed. The exploded logit model will be denoted as the rank model throughout the paper.

Alternatively, if the respondent evaluates the alternatives as a sequence of best then worst choices, either because they are constrained to do so, or because they find such an evaluation cognitively less burdensome, the same implied ranking can be expressed as a sequence of best then worst choices from the alternatives that remain:

$$P(A,B,C,D,E) = P_{Best}(A|A,B,C,D,E).P_{Worst}(E|B,C,D,E).P_{Best}(B|B,C,D).P_{Worst}(D|C,D) \quad (3)$$

However, such a formulation requires an expression for the probability of an alternative being the worst in a choice set. Marley and Louviere (2005) suggest that where the probability of an alternative being the best in a choice set of alternatives (say A in alternatives A, .., E, as before) is

$$P_{Best}(A) = \frac{e^{\beta x_A}}{\sum_{j=A,B,C,D,E} e^{\beta x_j}} , \quad (4)$$

the probability of an alternative being worst (say E from B,C,D,E) might be

$$P_{Worst}(E) = \frac{e^{-\beta x_E}}{\sum_{j=B,C,D,E} e^{-\beta x_j}} . \quad (5)$$

That is, the sign of the scale is reversed. Using this approach, and assuming the same preference order as for the rank model example, Equation (3) can be rewritten as

$$\text{P(A,B,C,D,E)} = \frac{e^{\mu_{B1}\beta x_A}}{\sum_{j=A,B,C,D,E} e^{\mu_{B1}\beta x_j}} \frac{e^{-\mu_{W1}\beta x_E}}{\sum_{j=B,C,D,E} e^{-\mu_{W1}\beta x_j}} \frac{e^{\mu_{B2}\beta x_B}}{\sum_{j=B,C,D} e^{\mu_{B2}\beta x_j}} \frac{e^{-\mu_{W2}\beta x_D}}{\sum_{j=C,D} e^{-\mu_{W2}\beta x_j}}. \quad (6)$$

The left hand side of Equation (6) implies that this is a ranking, however it is actually a probability that A is the first best (denoted B1) alternative (i.e., rank=1), E the first worst (W1) alternative (rank=5), B the second best (B2) alternative (rank=2), and D the second worst (W2) alternative (rank=4), with C having an implied rank of 3. Again, scale differences across the pseudo-observations can be identified, provided at least one μ_r is normalised, with the remaining scale parameters estimated relative to this value. The scales are labelled with the best-worst notation, to make clear that they are associated with one of the best-worst choices, not a rank. This best-worst-best-worst model will be denoted as the BWBW model throughout the paper.

In this paper, we will also investigate the possibility that individuals chose alternatives in the order of first best, second best, first worst, second worst. This model, which we will refer to as the BBWW model, can be represented as

$$\text{P(A,B,C,D,E)} = \frac{e^{\mu_{B1}\beta x_A}}{\sum_{j=A,B,C,D,E} e^{\mu_{B1}\beta x_j}} \frac{e^{\mu_{B2}\beta x_B}}{\sum_{j=B,C,D,E} e^{\mu_{B2}\beta x_j}} \frac{e^{-\mu_{W1}\beta x_E}}{\sum_{j=C,D,E} e^{-\mu_{W1}\beta x_j}} \frac{e^{-\mu_{W2}\beta x_D}}{\sum_{j=C,D} e^{-\mu_{W2}\beta x_j}}. \quad (7)$$

Several methods have been proposed for identifying differences in scale across the pseudo-observations of exploded logit models, including an application of the nested logit trick (Bradley and Daly 1994), and a direct multiplication of the utility by $\mu = e^{(\delta w)}$, where *w* are dummies for each of the ranks (Scarpa et al. 2009). These approaches treat any difference in scale as purely deterministic, however. An alternative approach is to use the scaled SMNL model, which is a specific case of the generalized multinomial logit (GMNL) model (Fiebig et al. 2010). In addition to allowing for scale to vary deterministically over alternatives or individuals, the SMNL model introduces a random disturbance into the scale component of utility, where this disturbance is fixed over all of an individual's observations. Thus the model can accommodate differing levels of engagement or ability to complete the choice task, without the necessity of finding suitable explanatory variables.

Let U_{nsj} denote the utility of alternative *j* perceived by respondent *n* in choice situation *s*. U_{nsj} may be partitioned into two separate components, an observed component of utility, V_{nsj} and a residual unobserved component, ε_{nsj}, such that

$$U_{nsj} = V_{nsj} + \varepsilon_{nsj}. \quad (8)$$

The observed component of utility is typically assumed to be a linear relationship of observed attribute levels, *x*, of each alternative *j* and their corresponding weights (parameters), β, such that

$$U_{nsj} = \mu_n \sum_{k=1}^{K} \beta_k x_{nsjk} + \varepsilon_{nsj}, \quad (9)$$

where β_k represents the marginal utility or parameter weight associated with attribute *k* for respondent *n* and the unobserved component, ε_{nsj}, is assumed to be independently and identically (IID) extreme value type 1 (EV1) distributed.

It is clear from Equations (9) that both μ_n and β_k cannot separately be estimated and as such, most discrete choice models make assumptions about μ_n such as $\mu_n = 1.0$, thus allowing β_k to be estimated. The utility specification in Equation (9) is flexible in that it allows for a number of different functional forms. The SMNL model assumes that different respondents have the same marginal utilities for each attribute being modelled but different error variances and hence scales, as shown in Equation (10).

$$\beta_{nk} = \mu_n \beta_k, \tag{10}$$

where

$$\mu_n = \mathrm{e}^{\left(\bar{\mu} + \sum_{q=1}^{Q} \delta_q w_q + \tau v_n\right)}. \tag{11}$$

$\bar{\mu}$ in Equation (11) denotes a mean parameter of scale, τ a variance parameter representing scale heterogeneity and v_n a draw from a standard Normal distribution representing the unobserved scale heterogeneity. Note that under this specification, scale is therefore assumed to be lognormally distributed. δ_q in Equation (11) represents parameters associated with covariates w_q which may be used to decompose the scale parameter.

In order for the model to be identified, it is necessary for some form of normalization to take place. This is done by setting the mean of μ_n to be one in the sample, which is accomplished by setting $\bar{\mu} = \frac{-\tau^2}{2}$.

Estimation of the model requires simulation of the log-likelihood over draws taken from z_n. The log-likelihood function of the model is

$$\log E(L) = \sum_{n=1}^{N} \log E\left(P_n^*\right). \tag{12a}$$

where

$$P_n^* = \prod_{s \in 1}^{S} \prod_{j \in 1}^{J} \left(P_{nsj}\right)^{y_{nsj}}. \tag{12b}$$

and where P_{nsj} are the choice probabilities calculated for the model and y_{nsj} equals one if alternative *j* is the chosen alternative in choice situation *s* shown to respondent *n*, and zero otherwise. Treatment of the log-likelihood function in this manner directly accounts for the panel nature of the SP data (see Revelt and Train 1998).

We have elected not to accommodate random preference heterogeneity in addition to random scale heterogeneity, using, for example, the generalised multinomial logit model (Fiebig et al. 2010), despite a cautionary note by Greene and Hensher (2010). Instead, preference heterogeneity is accommodated systematically by specifying attributes as differences between

the design levels and the respondents' self reported levels, as described in the next section. This approach yields highly significant improvements in model fit, and describes preference heterogeneity in a much more meaningful way than with purely random tastes. Random parameters could still be introduced to account for further preference heterogeneity, however such a treatment would likely need to be selective, as the representative utility is parameterised with 31 parameters.

In this paper, we test and compare, on the same dataset, the rank, BWBW and BBWW models. In addition, we test two specifications of the panel. In the conventional specification, all observations, irrespective of which pseudo-observation or rank they belong to, are treated as belonging to an individual. That is, draws v_n will vary between respondents only. In the second specification, the draws v_{nr} will vary both between respondent and rank *r*. This alternative approach might be valid, if individuals exhibit some random disturbance in scale between the ranks that extends beyond the systematic, population level difference that is also estimated. The validity of this alternative approach will be tested empirically.

3. Empirical setting

The empirical setting for the study is a stated preference survey focusing on preferences in an internet dating context. Hypothetical profiles of people were presented, with respondents required to evaluate each profile in terms of whether they would contact the person. Each profile was described in terms of five attributes: drinking habit (non drinker, casual drinker, moderate drinker), smoking habit (non smoker, ex smoker, current smoker), whether they have children (none currently, single parent, doesn't want children), their job type (unemployed, blue collar, white collar), and looks (below average, average, above average). Additionally, in the interests of obtaining willingness to pay measures, each profile listed a cost for contacting each individual (with potential levels of AUD10, AUD15 and AUD20), the level of which was not correlated with any other attribute level. Respondents were presented with nine scenarios, each of which contained five profiles. Figure 4 shows one such scenario, and details the wording used, the response mechanisms, and the combinations of levels used in one of the scenarios presented. Respondents were required to indicate the overall best and worst profiles (by selecting "best" and "worst" in the popup menus), as well as the best and worst profiles from the remaining three. No restrictions were placed on the order in which the best-worst decisions could be made. Consequently, it was possible to choose the first then second best profiles before choosing the two worst profiles. The actual sequence in which the best and worst options were selected was not recorded.

It must be acknowledged that the choice scenarios are somewhat different from real internet dating choice contexts, both in terms of the choice mechanisms, and the attributes used to describe the profiles. In particular, real profiles typically contain much more information including open ended text, job categories are typically much more extensive, and the looks of the candidate are typically conveyed through either a profile picture, or indirectly through an attribute such as 'body type'. Nonetheless, the profiles essentially present a simplification of a real profile from a dating site.

The data were collected from an internet panel (The Online Research Unit http://www.theoru.com/) in June 2010. Twelve hundred and seven respondents were sampled from a potential panel of 300,000 with only currently single individuals eligible to participate in the survey. As part of the study, respondents were randomly assigned to one of five different experimental designs involving having to answer nine complete best worst choice tasks. In addition to the choice questions, numerous other questions were asked once the best-worst scenarios were completed. Crucially, respondents were asked to describe themselves on each of the non-cost dimensions used to describe the profiles. Obviously respondents might answer such questions with some degree of bias (a disproportionate number did not describe themselves as of below average looks, for example). Nonetheless, such responses provide some indication of

what they are like relative to the profiles, and we will see that this has important implications for the modelling of the choices. Other socio-demographic information was also collected, including gender, age, hours of work per week, income, and whether they have children. In this paper, only male respondents are used for analysis, as notable differences were found in the dating preferences of the two genders. Table 1 outlines the socio-demographic profile of the 461 male respondents.

Table 1: Socio-demographic profile of 461 male respondents

Age	**24 or under**	**25 to 34**	**35 to 44**	**45 to 54**	**55 to 64**	**65 and over**	**Won't say**
	37	54	70	102	131	64	3
Income ($'000)	**Under 10**	**10-15**	**15-20**	**20-30**	**30-40**	**40-50**	
	23	36	58	51	46	48	
	50-60	**60-80**	**80-100**	**100-120**	**Over 120**	**Won't say**	
	39	46	27	16	20	51	
Drinking habit	**Non drinker**	**Casual drinker**	**Moderate drinker**				
	92	243	126				
Smoking habit	**Non smoker**	**Ex smoker**	**Smoker**				
	250	112	99				
Children	**None currently**	**Single parent**	**Don't want children**				
	259	80	122				
Job	**Unemployed**	**Blue collar**	**White collar**				
	133	102	226				
Looks	**Below average**	**Average**	**Above average**				
	32	367	62				

Rather than estimate models directly using the design attributes, we transform the data for the non cost attributes so as to estimate models based on differences between the design attributes and the respondents' self reported level for each attribute. The data transformation was done by creating dummy codes for differences between the respondents' reported level and the design attribute they were shown. An example of the data structure used in the analysis is presented in Table 2 for the drinking habit attribute. Note that the table represents the coding structure for the three types of respondents – those who don't drink, those who drink a moderate amount, and those who drink casually. Other non cost attributes were similarly transformed. The cost attribute was entered into the model using the values shown to respondents.

Interpretation of the model outputs derived from using the data transformation in this manner requires some explanation. The parameter estimates from the reported models should be interpreted as representing the sensitivity to differences between the respondent's self reported level of that attribute, and the attribute of the prospective contact.

Table 2: Example dummy coding used for analysis (drinking habit)

		Coding in data					
Respondent	**Design attribute**	**D-C**	**D-M**	**C-D**	**C-M**	**M-D**	**M-C**
Don't (D)	**Casual (C)**	1	0	0	0	0	0
Don't (D)	**Moderate (M)**	0	1	0	0	0	0
Don't (D)	**Don't (D)**	0	0	0	0	0	0
Casual (C)	**Don't (D)**	0	0	1	0	0	0
Casual (C)	**Moderate (M)**	0	0	0	1	0	0
Casual (C)	**Casual (C)**	0	0	0	0	0	0
Moderate (M)	**Don't (D)**	0	0	0	0	1	0
Moderate (M)	**Casual (C)**	0	0	0	0	0	1
Moderate (M)	**Moderate (M)**	0	0	0	0	0	0

4. Results

4.1 Rank model

Table 3 presents the results for an MNL model and two SMNL models assuming ranked data estimated using the conventional exploded logit formulation (Equation 2). The two SMNL models, both estimated using 200 Halton draws, differ in terms of how the panel structure of the data is treated. In the first SMNL model, scale is assumed to be constant within each ranking of each individual (i.e., each different pseudo-observation is assumed to have a different scale within each individual; SMNL, panel = 9). In the second SMNL model, scale is assumed to be constant within each individual independent of the ranking of the pseudo-observation being modelled (SMNL, panel = 36). In all three models, the cost parameter is insignificant, suggesting that respondents are largely indifferent to the cost to contact (at least within the range of levels considered in the experiment). Given that the cost parameter is not statistically significant, we are unable to compute willingness to pay measures. The table is useful in illustrating the structure of the models and the outputs produced. For example, six sensitivities are retrieved for the smoking habit attribute, two each for non smokers, ex smokers and smokers. Non smokers exhibit a clear preference against current smokers, with a much milder preference against ex smokers. By contrast, smokers prefer other smokers to ex and non smokers, however their disutility associated with the latter is far less than that the non smokers associate with smokers. Drawing on the SMNL (panel=36) results from Table 2, it can further be seen that ex smokers have a very mild preference for non smokers over other ex smokers, and a strong preference against smokers, but not as strong as for non smokers. Clear differences across the sensitivities can be observed for all non cost attributes, with the results presenting interesting findings on the human condition! For example, those with below average looks certainly have a preference for those with above average looks, but their utility is far outweighed by the magnitude (and significance) of the disutility those with above average looks associate with those of below average looks. The introduction of these interactions led to a dramatic improvement in model fit from the main effects only model.

Examination of the MNL model results show that several parameters are statistically insignificant, including M-C (drinking moderate vs causal), E-N (ex smokers vs non-smokers), Sp-N (single parents vs no children contact), and B-W (blue collar worker vs white collar contact) which signify, respectively, that moderate drinkers are indifferent between moderate and casual drinkers, ex smokers are indifferent between ex and non smokers, single parents are indifferent between other single parents and those who have no children, and blue collar workers are indifferent between blue and white collar workers. Both the magnitudes of the significant parameters and the set of insignificant parameters are plausible.

The second and third models introduce both across rank scale heterogeneity and between respondent scale heterogeneity through the SMNL model. The second model handles the panel nature of the data by treating each combination of individual and rank as a single individual in the dataset. That is, the random parameter in the scale will remain invariant across all nine observations for each rank for each individual, but may vary across ranks for that individual. In contrast, for the third model, the random parameter remains invariant across all 36 observations, and does not vary by rank. While SMNL heterogeneity models provide a highly significant improvement on the MNL model, the third model fits the data better than the second, with a reduction in log likelihood of over 200 units, with no increase in number of parameters. This suggests that varying the stochastic disturbance by rank instead of by individual is not advisable. The tau parameter, which controls the extent of stochastic scale heterogeneity, is highly significant in both models, suggesting that respondents exhibit different amounts of error variance.

Table 3: Models with ranked data

				MNL		SMNL, panel=9		SMNL, panel=36	
	Trait	**Them**	**Date**	**Param.**	(***t*-ratio**)	**Param.**	(***t*-ratio**)	**Param.**	(***t*-ratio**)
D-C	**Drinking habit**	**Don't**	**Casual**	-0.240	(-4.49)	-0.596	(-8.01)	-0.576	(-7.06)
D-M		**Don't**	**Moderate**	-0.445	(-8.23)	-1.007	(-11.92)	-1.040	(-13.78)
C-D		**Casual**	**Don't**	-0.203	(-6.24)	-0.286	(-5.63)	-0.243	(-5.25)
C-M		**Casual**	**Moderate**	-0.164	(-5.07)	-0.271	(-5.10)	-0.236	(-4.68)
M-D		**Moderate**	**Don't**	-0.525	(-11.56)	-0.935	(-12.39)	-0.847	(-12.60)
M-C		**Moderate**	**Casual**	0.078	(1.75)	0.111	(1.63)	0.080	(1.25)
N-E	**Smoking habit**	**Non smoker**	**Ex smoker**	-0.396	(-12.56)	-0.763	(-14.33)	-0.792	(-14.58)
N-S		**Non smoker**	**Smoker**	-1.902	(-47.49)	-4.411	(-18.22)	-4.199	(-17.22)
E-N		**Ex smoker**	**Non smoker**	0.041	(0.90)	0.141	(2.38)	0.156	(2.76)
E-S		**Ex smoker**	**Smoker**	-1.498	(-27.54)	-3.429	(-16.55)	-3.184	(-16.33)
S-N		**Smoker**	**Non smoker**	-0.370	(-7.38)	-0.737	(-10.63)	-0.597	(-10.01)
S-E		**Smoker**	**Ex smoker**	-0.404	(-8.05)	-0.733	(-10.44)	-0.606	(-10.49)
N-Sp	**Children**	**None**	**Single parent**	-0.322	(-10.25)	-0.582	(-10.88)	-0.519	(-10.54)
N-Dw		**None**	**Don't want**	-0.297	(-8.68)	-0.429	(-8.18)	-0.376	(-8.12)
Sp_N		**Single parent**	**None**	0.071	(1.28)	0.117	(1.35)	0.156	(2.02)
Sp_Dw		**Single parent**	**Don't want**	-0.263	(-4.50)	-0.417	(-5.07)	-0.275	(-4.68)
Dw-N		**Don't want**	**None**	-0.202	(-4.29)	-0.382	(-5.12)	-0.345	(-5.10)
Dw-Sp		**Don't want**	**Single parent**	-0.744	(-15.49)	-1.275	(-13.85)	-1.054	(-13.26)
U-B	**Job**	**Unemployed**	**Blue collar**	0.387	(8.81)	0.564	(7.30)	0.514	(7.61)
U-W		**Unemployed**	**White collar**	0.421	(9.26)	0.653	(8.34)	0.560	(8.10)
B-U		**Blue collar**	**Unemployed**	-0.596	(-11.91)	-1.013	(-12.43)	-0.899	(-12.22)
B-W		**Blue collar**	**White collar**	-0.062	(-1.25)	-0.110	(-1.61)	-0.063	(-0.97)
W-U		**White collar**	**Unemployed**	-0.968	(-26.15)	-1.679	(-16.55)	-1.515	(-15.80)
W-B		**White collar**	**Blue collar**	-0.283	(-8.22)	-0.521	(-9.34)	-0.464	(-9.39)
Ba-A	**Looks**	**Below avg.**	**Average**	0.202	(2.26)	0.471	(3.46)	0.326	(2.44)
Ba-Aa		**Below avg.**	**Above avg.**	0.451	(5.02)	0.702	(5.18)	0.533	(5.71)
A-Ba		**Average**	**Below avg.**	-0.655	(-24.38)	-1.266	(-17.26)	-1.188	(-16.11)
A-Aa		**Average**	**Above avg.**	0.301	(11.01)	0.544	(11.72)	0.534	(11.98)
Aa-Ba		**Above avg.**	**Below avg.**	-1.592	(-22.05)	-3.669	(-16.34)	-3.590	(-16.47)
Aa-A		**Above avg.**	**Average**	-0.745	(-11.27)	-1.606	(-13.53)	-1.606	(-12.71)
COST	**Cost to contact ($10,$15,$20)**			-0.004	(-1.52)	-0.001	(-0.26)	-0.001	(-0.24)
Tau						0.752	(28.04)	0.713	(23.04)
Rank 2						-0.372	(-5.44)	-0.340	(-9.13)
Rank 3						-0.616	(-9.14)	-0.600	(-15.71)
Rank 4						-1.035	(-13.68)	-0.893	(-16.90)
LL(0)				-19863.3		-19863.3		-19863.3	
LL(beta)				-16202.1		-15609.9		-15401.4	
Number of parameters				31		35		35	
ρ^2				0.184		0.214		0.225	
Adjusted ρ^2				0.183		0.212		0.223	
AIC				1.956		1.885		1.860	
Observations				16596		16596		16596	
Respondents				461		461		461	

The inclusion in the scale of covariates for each rank's choice suggests that scale is lower for choices of lower rank, with scale monotonically decreasing in both SMNL models. Such a finding is consistent with most of the literature, as detailed earlier. Of note, however, is Scarpa et al. (2009), who employed the exploded logit method with the pseudo-observations representing sequential best choice, but with a best-worst response format. They found that the worst choice, or equivalently the ranking of the fourth (and consequently fifth) best alternative from the five alternatives in the choice task, had greater scale than ranks two and three. This

suggests that scale decreases as the number of alternatives decreases in the choice task (but not the pseudo-observations used for modelling).

A comparison of the SMNL and MNL parameter estimates suggests that differences between the two are not merely a consequence of a different magnitude in mean scale. That is, there is not a consistent ratio between the corresponding parameters between the two models, and so different sensitivities are revealed. This might be due either to scale heterogeneity across individuals or ranks. Model results not reported here suggest that both play a role in the differences (through the estimation of the model without scale covariates and the random disturbance tau, respectively). Of note is parameter E-N (ex smoker vs none smoker contact), which is insignificant in the MNL model yet significant in both SMNL models; and Sp-N (single parents vs no children contact), which is significant at the 90 percent confidence level in the better fitting SMNL model, but insignificant in the other two models. So, in addition to differing sensitivities and a vastly better model fit, the third model identifies two more parameters than the MNL model as significant.

4.2 Best-worst-best-worst (BWBW) model

Table 4 presents models that treat the choice process as a sequence of best then worst choices. As with the rank model, the SMNL specification is a significant improvement on the MNL specification, with the panel specification across all 36 of an individual's observations leading to a much better model fit than across the nine observations from each best or worst decision. Consequently, this model is accepted as the preferred BWBW model. The tau parameter is nearly identical to the rank model, both in magnitude and significance, suggesting a similar level of inter-respondent scale heterogeneity.

While caution is warranted when comparing the log likelihoods of the models, due to differences in the treatment of the pseudo-observations, the BWBW model has log likelihoods and ρ^2 values that are slightly larger than the rank model, with a slightly smaller AIC. Such a finding provides tentative evidence that the respondents are employing a BWBW process. This seems plausible, since this is what they were instructed to do, without being forced.

The scale covariates in the BWBW model tell a different story to the rank model. Relative to the first best choice, the first worst choice has the next lowest scale (with a parameter in the preferred SMNL model of -0.435), followed by the second best choice (-0.491), then the second worst choice (-0.855). Rather than being consistent with the decrease in scale of the rank model, the scale decreases with each successive choice implied by the estimated model. The model itself relies on a specific handling of the data, with each successive best then worst choice having one less alternative in the pseudo-observation. Indeed, the consistency of the scale covariates with the rank model seemingly lies in a decreasing scale as the number of alternatives *in the model* decreases. It is not apparent why this might be the case, although it provides a clue that perhaps the decrease in scale is more of an artefact of the modelling processes employed than a behavioural phenomenon.

4.3 Best-best-worst-worst (BBWW) model

Table 5 presents the models that treat the choice process as a sequence of two best then two worst choices. Again, the SMNL specification is a significant improvement on the MNL specification, with the panel specification across all 36 of an individual's observations leading to a much better model fit than across the nine observations from each best or worst decision. Further, tau is nearly identical to both the rank and BWBW model, both in magnitude and significance, suggesting a similar level of inter-respondent scale heterogeneity across all three models. The log likelihood and ρ^2 of the BBWW model sits in between the rank and BWBW models.

Table 4: Models with best-worst-best-worst data

				MNL		SMNL, panel=9		SMNL, panel=36	
	Trait	**Them**	**Date**	**Param.**	(***t*-ratio**)	**Param.**	(***t*-ratio**)	**Param.**	(***t*-ratio**)
D-C	**Drinking habit**	**Don't**	**Casual**	-0.297	(-5.51)	-0.689	(-9.00)	-0.627	(-7.73)
D-M		**Don't**	**Moderate**	-0.475	(-8.77)	-1.070	(-12.05)	-1.049	(-13.82)
C-D		**Casual**	**Don't**	-0.193	(-5.88)	-0.274	(-5.41)	-0.236	(-5.01)
C-M		**Casual**	**Moderate**	-0.170	(-5.23)	-0.299	(-5.50)	-0.258	(-4.98)
M-D		**Moderate**	**Don't**	-0.497	(-10.90)	-0.926	(-12.42)	-0.842	(-12.59)
M-C		**Moderate**	**Casual**	0.089	(1.97)	0.114	(1.67)	0.078	(1.23)
N-E	**Smoking habit**	**Non smoker**	**Ex smoker**	-0.418	(-13.12)	-0.811	(-14.36)	-0.825	(-14.82)
N-S		**Non smoker**	**Smoker**	-1.930	(-47.89)	-4.424	(-18.21)	-4.269	(-17.40)
E-N		**Ex smoker**	**Non smoker**	0.086	(1.85)	0.193	(3.04)	0.177	(2.97)
E-S		**Ex smoker**	**Smoker**	-1.456	(-27.32)	-3.331	(-16.45)	-3.146	(-16.47)
S-N		**Smoker**	**Non smoker**	-0.453	(-8.89)	-0.848	(-11.60)	-0.643	(-10.72)
S-E		**Smoker**	**Ex smoker**	-0.442	(-8.80)	-0.796	(-10.96)	-0.623	(-10.81)
N-Sp	**Children**	**None**	**Single parent**	-0.323	(-10.23)	-0.590	(-10.91)	-0.524	(-10.61)
N-Dw		**None**	**Don't want**	-0.307	(-8.91)	-0.424	(-8.12)	-0.372	(-7.99)
Sp_N		**Single parent**	**None**	0.064	(1.13)	0.135	(1.57)	0.171	(2.21)
Sp_Dw		**Single parent**	**Don't want**	-0.273	(-4.70)	-0.405	(-4.99)	-0.275	(-4.67)
Dw-N		**Don't want**	**None**	-0.218	(-4.58)	-0.393	(-5.29)	-0.343	(-5.06)
Dw-Sp		**Don't want**	**Single parent**	-0.746	(-15.57)	-1.190	(-13.30)	-1.004	(-13.02)
U-B	**Job**	**Unemployed**	**Blue collar**	0.376	(8.57)	0.533	(6.91)	0.495	(7.32)
U-W		**Unemployed**	**White collar**	0.392	(8.65)	0.601	(7.67)	0.523	(7.71)
B-U		**Blue collar**	**Unemployed**	-0.577	(-11.53)	-1.006	(-12.05)	-0.886	(-12.00)
B-W		**Blue collar**	**White collar**	-0.067	(-1.33)	-0.120	(-1.68)	-0.077	(-1.16)
W-U		**White collar**	**Unemployed**	-0.980	(-26.46)	-1.689	(-16.53)	-1.514	(-15.91)
W-B		**White collar**	**Blue collar**	-0.308	(-8.89)	-0.537	(-9.42)	-0.462	(-9.23)
Ba-A	**Looks**	**Below avg.**	**Average**	0.218	(2.45)	0.352	(2.61)	0.263	(2.06)
Ba-Aa		**Below avg.**	**Above avg.**	0.410	(4.61)	0.589	(4.43)	0.468	(5.13)
A-Ba		**Average**	**Below avg.**	-0.646	(-24.34)	-1.212	(-16.96)	-1.160	(-16.33)
A-Aa		**Average**	**Above avg.**	0.345	(12.48)	0.581	(12.03)	0.549	(12.05)
Aa-Ba		**Above avg.**	**Below avg.**	-1.684	(-23.05)	-3.791	(-16.56)	-3.679	(-16.77)
Aa-A		**Above avg.**	**Average**	-0.799	(-11.86)	-1.718	(-13.70)	-1.658	(-12.52)
COST	**Cost to contact ($10,$15,$20)**			-0.003	(-0.97)	-0.002	(-0.59)	-0.003	(-0.75)
Tau						0.748	(27.83)	0.714	(23.13)
Best 2						-0.562	(-7.95)	-0.491	(-11.90)
Worst 1						-0.469	(-7.27)	-0.435	(-12.11)
Worst 2						-0.911	(-12.04)	-0.855	(-17.15)
LL(0)				-19863.3		-19863.3		-19863.3	
LL(beta)				-16139.0		-15589.0		-15385.2	
Number of parameters				31		35		35	
ρ^2				0.188		0.215		0.225	
Adjusted ρ^2				0.186		0.214		0.224	
AIC				1.949		1.883		1.858	
Observations				16596		16596		16596	
Respondents				461		461		461	

Table 5: Models with best-best-worst-worst data

				MNL		SMNL, panel=9		SMNL, panel=36	
	Trait	**Them**	**Date**	**Param.**	(***t*-ratio**)	**Param.**	(***t*-ratio**)	**Param.**	(***t*-ratio**)
D-C	**Drinking habit**	**Don't**	**Casual**	-0.273	(-5.08)	-0.641	(-8.54)	-0.603	(-7.39)
D-M		**Don't**	**Moderate**	-0.464	(-8.60)	-1.035	(-12.07)	-1.039	(-13.79)
C-D		**Casual**	**Don't**	-0.197	(-6.02)	-0.280	(-5.54)	-0.236	(-5.05)
C-M		**Casual**	**Moderate**	-0.169	(-5.20)	-0.285	(-5.33)	-0.247	(-4.80)
M-D		**Moderate**	**Don't**	-0.509	(-11.17)	-0.931	(-12.46)	-0.849	(-12.63)
M-C		**Moderate**	**Casual**	0.087	(1.95)	0.097	(1.43)	0.076	(1.20)
N-E	**Smoking habit**	**Non smoker**	**Ex smoker**	-0.403	(-12.66)	-0.780	(-14.43)	-0.801	(-14.73)
N-S		**Non smoker**	**Smoker**	-1.914	(-47.57)	-4.447	(-18.19)	-4.260	(-17.29)
E-N		**Ex smoker**	**Non smoker**	0.064	(1.37)	0.168	(2.78)	0.167	(2.91)
E-S		**Ex smoker**	**Smoker**	-1.463	(-27.35)	-3.379	(-16.40)	-3.178	(-16.39)
S-N		**Smoker**	**Non smoker**	-0.413	(-8.11)	-0.795	(-11.21)	-0.614	(-10.28)
S-E		**Smoker**	**Ex smoker**	-0.409	(-8.17)	-0.758	(-10.68)	-0.604	(-10.50)
N-Sp	**Children**	**None**	**Single parent**	-0.324	(-10.28)	-0.587	(-11.00)	-0.522	(-10.60)
N-Dw		**None**	**Don't want**	-0.303	(-8.81)	-0.410	(-7.96)	-0.364	(-7.88)
Sp_N		**Single parent**	**None**	0.067	(1.19)	0.132	(1.53)	0.166	(2.14)
Sp_Dw		**Single parent**	**Don't want**	-0.258	(-4.44)	-0.397	(-4.87)	-0.270	(-4.63)
Dw-N		**Don't want**	**None**	-0.208	(-4.38)	-0.393	(-5.29)	-0.345	(-5.12)
Dw-Sp		**Don't want**	**Single parent**	-0.746	(-15.60)	-1.230	(-13.71)	-1.027	(-13.16)
U-B	**Job**	**Unemployed**	**Blue collar**	0.395	(9.01)	0.559	(7.33)	0.510	(7.52)
U-W		**Unemployed**	**White collar**	0.412	(9.05)	0.636	(8.18)	0.544	(7.96)
B-U		**Blue collar**	**Unemployed**	-0.582	(-11.65)	-1.009	(-12.30)	-0.894	(-12.17)
B-W		**Blue collar**	**White collar**	-0.070	(-1.40)	-0.116	(-1.69)	-0.066	(-1.00)
W-U		**White collar**	**Unemployed**	-0.982	(-26.52)	-1.688	(-16.61)	-1.515	(-15.84)
W-B		**White collar**	**Blue collar**	-0.301	(-8.69)	-0.528	(-9.43)	-0.460	(-9.32)
Ba-A	**Looks**	**Below avg.**	**Average**	0.194	(2.19)	0.415	(3.06)	0.296	(2.29)
Ba-Aa		**Below avg.**	**Above avg.**	0.411	(4.60)	0.663	(4.97)	0.512	(5.56)
A-Ba		**Average**	**Below avg.**	-0.645	(-24.24)	-1.238	(-17.15)	-1.175	(-16.16)
A-Aa		**Average**	**Above avg.**	0.320	(11.60)	0.560	(11.91)	0.542	(12.11)
Aa-Ba		**Above avg.**	**Below avg.**	-1.636	(-22.42)	-3.740	(-16.28)	-3.644	(-16.57)
Aa-A		**Above avg.**	**Average**	-0.766	(-11.46)	-1.651	(-13.59)	-1.634	(-12.61)
COST	**Cost to contact ($10,$15,$20)**			-0.003	(-1.24)	-0.002	(-0.48)	-0.003	(-0.82)
Tau						0.750	(27.99)	0.716	(23.14)
Best 2						-0.375	(-5.50)	-0.344	(-9.17)
Worst 1						-0.681	(-10.33)	-0.618	(-15.61)
Worst 2						-0.914	(-12.09)	-0.858	(-17.22)
LL(0)				-19863.3		-19863.3		-19863.3	
LL(beta)				-16173.8		-15606.9		-15395.2	
Number of parameters				31		35		35	
ρ^2				0.186		0.214		0.225	
Adjusted ρ^2				0.184		0.213		0.223	
AIC				1.953		1.885		1.860	
Observations				16596		16596		16596	
Respondents				461		461		461	

The scale covariates in the BBWW model decrease over the choice order that the model implies, with the second best model having a covariate of -0.344, followed by the first worst choice (-0.618), then the second worst choice (-0.858). The correlation once again is with the size of the choice set in each pseudo-observation.

4.4 Comparison of parameter estimates across data structures

Thus far, the model results suggest that the tau parameter is highly consistent across the rank, BWBW, and BBWW models, thus suggesting that the models are detecting similar degrees of scale heterogeneity. Of similar interest however is whether the parameter estimates across models estimated on different data structures are consistent. Despite the explicit handling of scale, both deterministically and stochastically, differences in scale may still remain across the datasets, precluding any direct comparison of parameter estimates. A convenient method often employed is to estimate willingness to pay values and their associated standard errors, thus obtaining measures that are free of scale. Unfortunately, the cost parameter is not significant in any of the models. As an alternative, for each model *m*, we estimate a complete set of marginal rates of substitution, MRS_m, between all *K* parameters in the model, together with their standard errors:

$$MRS_m = \{\beta_{k_1} / \beta_{k_2} \mid k_1 \neq k_2\}; \forall k_1, k_2 \in K$$

$$SE_m = \{s.e.(\beta_{k_1} / \beta_{k_2}) \mid k_1 \neq k_2\}; \forall k_1, k_2 \in K$$

For each marginal rate of substitution $mrs \in MRS$, we then test for statistical equivalence between the three pairs of models that can be formed from the rank, BWBW, and BBWW models. While some of the comparisons cannot be made due to insignificance in one or both of the mrs_m, all remaining mrs are found to be statistically equivalent across all three models. Specifically, for the rank to BWBW comparison, 728 mrs are statistically equivalent, while for 202 at least one mrs_m is insignificant. For the rank to BBWW comparison, the numbers are 738 and 192 respectively, and for BWBW to BBWW, the numbers are also 738 and 192 respectively. While the three models differ in the numbers of parameters and marginal rates of substitution that are significant within the model, where a comparison can be made, the marginal rates of substitution are equal. Hence we can claim that the sensitivities are the same, irrespective of the way in which the model is specified. This is a useful finding, as it suggests that even if the respondent adopts a choice sequence that does not match the model that is applied, asymptotically the same sensitivities will be retrieved. Of course this is one empirical finding, which would need further validation on other datasets before any generalisations can be made.

An alternative approach to evaluate the performance of different data structures is to compare the predicted choice probabilities for each alternative with the observed counts for each alternative across the sample. Table 6 presents for the three data structures, the true sample shares for the highest preferred alternatives combined with the estimated market shares based on the SMNL (panel 36) model for each alternative. The SMNL (panel 36) model was selected as this model appears to have performed the best on all three data structures. In bold are the percentages that are closest to the true data shares for each alternative. Although only minor differences exist, the SMNL model based on the rank exploded data reproduces the market shares for alternatives A and D better than the other data structures, whilst the BWBW SMNL model reproduces the market share for alternative B better than models applied to other data structures. The SMNL model applied to the BBWW data structure best reproduces the market shares for alternatives C and D when compared to the SMNL models applied to the other data structures. Looking at the deviation from the true market share, the SMNL model applied to the rank exploded data tends to produce smaller differences on average than the other data structures, with the BWBW the largest average deviation from the true market shares. Whilst i) not a formal statistical test, and ii) only small differences are observed in the modelled choice shares, the above provides some limited evidence that the model applied to the rank exploded data provides a better description of the observed outcomes than the other two data structures.

Table 6: True shares versus predicted shares

		Alt A	Alt B	Alt C	Alt D	Alt E
Data structure	**True shares**	0.21138	0.19788	0.19884	0.19667	0.19523
Rank	Estimated share	**0.20453**	0.19865	0.19642	**0.20195**	0.19843
	Dev. from true share	**0.00684**	0.00078	0.00242	**0.00528**	0.00321
BWBW	Estimated share	0.20443	**0.19852**	0.19638	0.20217	0.19850
	Dev. from true share	0.00695	**0.00064**	0.00246	0.00549	0.00327
BBWW	Estimated share	0.20443	0.19860	**0.19649**	0.20206	**0.19843**
	Dev. from true share	0.00695	0.00072	**0.00236**	0.00539	**0.00320**

4.5 Separate models for each pseudo-observation

The analysis this far has assumed that the sensitivities to the attributes are consistent over each of the pseudo-observations. However, estimation of separate models for each pseudo-observation, be they ranks or best or worst choices in a form of best-worst choice sequence, calls into question this assumption. We present, in figures 5 and 6, plots of the parameters estimated on each of the pseudo-observations, relative to the first choice (i.e., the first best choice, in all the models we have examined). All models were estimated using the SMNL model with 200 Halton draws, with the same parameters in the representative utility as with models previously reported in the paper, and no scale covariates. Pairs of parameters are only plotted if both parameter values are statistically significant at the 95 percent confidence level. Further, to condense the plots, some of the parameter pairs are scaled by a constant, including by a negative number where the parameter estimates are positive.

Figure 5 presents the parameter plots for the rank data. Of the 31 parameters estimated in the representative utility, four parameters are omitted as they are not significant for the first rank. Numerous parameters were then omitted from the plots for each rank, due to being insignificant at that rank; notably, two further parameters at rank two, 12 parameters at rank three, and 15 parameters at rank four. An increase in the number of insignificant parameters with each rank is not surprising, given the decrease in scale over the ranks. However, while more rigorous testing is necessary to draw definitive conclusions, the lack of a clear trend line for each of the ranks over the remaining parameters suggests that the differences in parameter values are unlikely to be due merely to differences in scale.

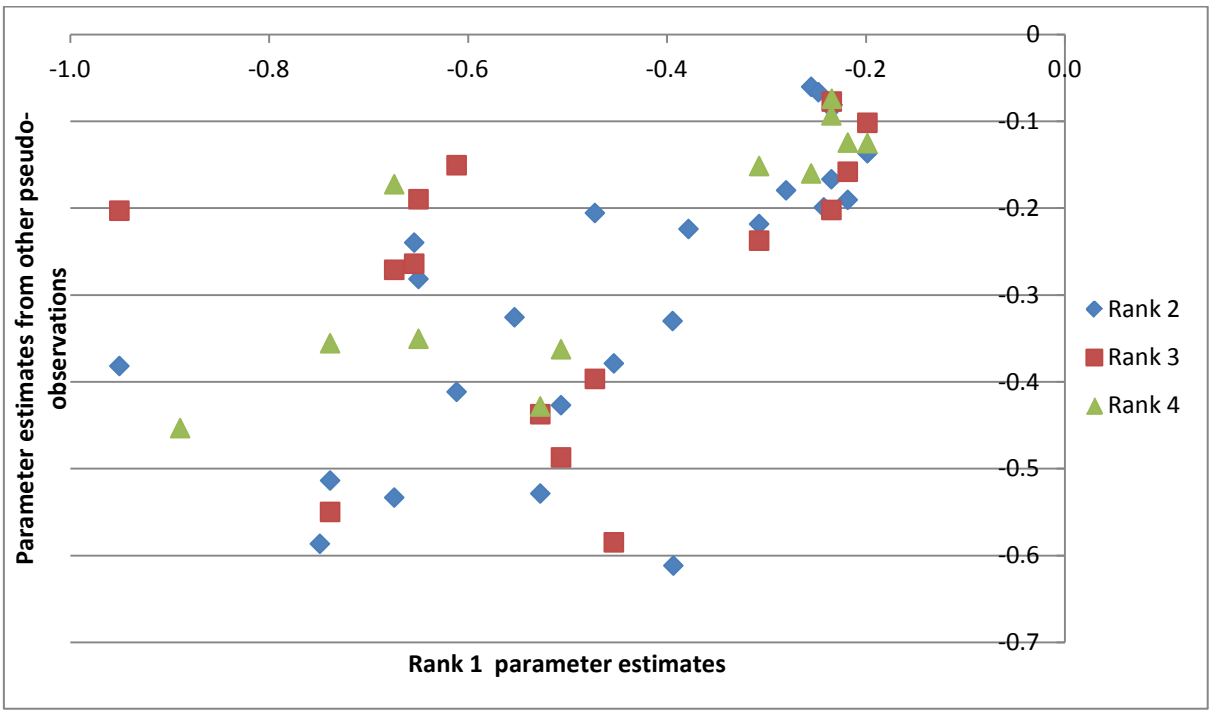

Fig 5: Parameter plots for rank models

Figure 6 presents the parameter plots for each of the pseudo-observations from the BWBW data. The same model is used for the first best choice as for rank one, and so as before, four parameters from 31 are dropped from the plots due to insignificance. Further parameters were omitted from each of the pseudo-observations due to insignificance: first worst choice (seven), second best choice (two), and second worst choice (16). Again, for each of the pseudo-observations, large deviations from a trend line are apparent, calling into question the consistency of the parameter estimates across the pseudo-observations.

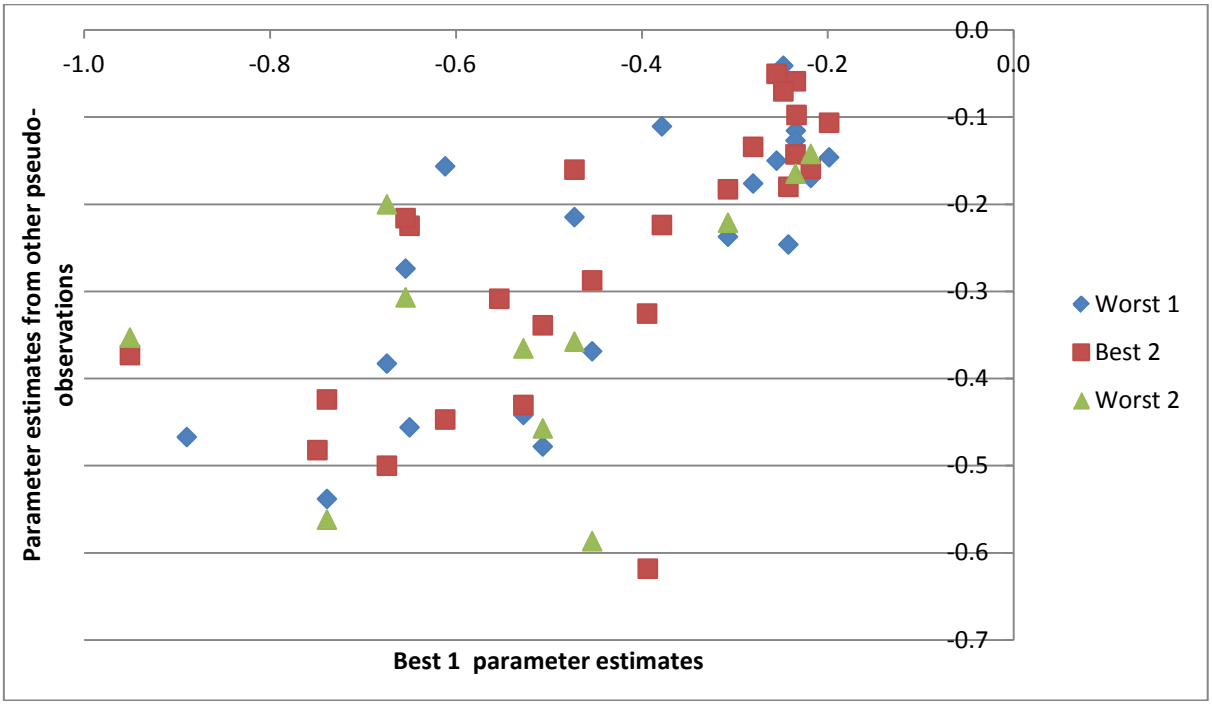

Fig 6: Parameter plots for BWBW models

5. Discussion and conclusion

In this paper, we have estimated models on ranked and best-worst data that demonstrate the importance of differences in scale both between rank, and between respondent. A conventional treatment of the panel, with the random component of scale varying across all of an individual's observations, outperforms an alternative specification, where the random component of scale varies across each rank, only, for each individual.

The models provide tentative evidence that, if the choices are framed as best-worst, but not enforced, the best-worst model outperforms the rank and BBWW models. However, the marginal rates of substitution, where significant, are consistent across each of the three models, suggesting that a misspecification of the model will not bias the parameter estimates.

While not forcing a particular response order allowed us to see which rank expansion method (rank, BWBW, BBWW) best fit the data, this approach cannot provide definitive conclusions about the merits of ranked choices verses best-worst choices. Indeed, it is possible that subsets of respondents within the sample are employing each method. Recording the order of completion might provide valuable insights here. An alternative would be to force different subsets to either rank or provide the best-worst responses, and compare the results of each. These remain areas for future research.

Scale heterogeneity across the individual is important, so long as we can be confident that the SMNL model is uncovering the correct form of heterogeneity. Alternatively, the SMNL model might be identifying random preference heterogeneity. Fiebig et al. (2010) find that scale heterogeneity has a stronger impact than preference heterogeneity across a number of datasets.

Nonetheless, this is an empirical question which the estimation of a GMNL model on this data would answer.

References

Ben-Akiva, M., Morikawa, T. and Shiroishi, F. (1991). Analysis of the reliability of preference ranking data. *Journal of Business Research*, 23(3), 253-268.

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.

Chapman, R. G. and Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*, 19(3), 288-301.Fiebig, D. G., Keane, M. P., Louviere, J. and Wasi, N. (2010). The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Science*, 29(3), 393-421.

Greene, W. H., & Hensher, D. A. (2010). Does scale heterogeneity across individuals matter? An empirical assessment of alternative logit models. *Transportation*, *10*(3), 413-428.

Hausman, J. and Ruud, P. (1987). Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics*, 34(1-2), 83-104.

Lancsar, E. and Louviere, J. (2008). Estimating individual level discrete choice models and welfare measures using best worst choice experiments and sequential best worst MNL. CenSoC Working Paper Series.

Louviere, J. J., Street, D., Burgess, L., Wasi, N., Islam, T., & Marley, A. A. J. (2008). Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. *Journal of Choice Modelling*, *1*(1), 128–163.

Luce, R. D. and Suppes, P. (1965). Preference, utility, and subjective probability. Handbook of mathematical psychology. Luce, R. D., Bush, R. R. and Galanter, E. New York, Wiley. III, 249-410.

Marley, A. and Louviere, J. (2005). Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology*, 49(6), 464-480.

Marley, A. A. J. and Pihlens, D. (2010). Models of Best-Worst Choice and Ranking Amoung, and in, Multiattribute Options (Profiles). CenSoC Working Paper Series.

Revelt, D. and Train, K. (1998). Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of Economics and Statistics*, 80(4), 647-657.

Scarpa, R., Notaro, S., Raffelli, R., Pihlens, D. and Louviere, J. (2009). Exploring scale effects of best/worst rank ordered choice data to estimate visitors' benefits from alpine transhumance. International Choice Modeling Conference.