Discrete choice models: scale heterogeneity and why it matters

Katrina J. Davis [a,b*], Michael Burton[b,c] and Marit E. Kragt[b]

[a]Centre of Excellence for Environmental Decisions, University of Queensland, St Lucia, Qld 4072, Australia
[b]School of Agricultural and Resource Economics, The University of Western Australia
[c]School of Social Sciences, University of Manchester

*E-mail address: k.davis@uq.edu.au

14 May 2016
Example Working Paper 1602
School of Agricultural and Resource Economics
http://www.are.uwa.edu.au

Citation: Davis, K.J., Burton, M. and Kragt, M.E. (2016) *Discrete choice models: scale heterogeneity and why it matters*, Working Paper 1602, School of Agricultural and Resource Economics, University of Western Australia, Crawley, Australia.

Discrete choice models: scale heterogeneity and why it matters

Katrina J. Davis, Michael Burton and Marit E. Kragt

Abstract: Models to analyse discrete choice data that account for heterogeneity in error variance (scale) across respondents are increasingly common, e.g. heteroscedastic conditional logit or scale adjusted latent class models. In this paper we do not question the need to allow for scale heterogeneity. Rather, we examine the interpretation of results from these models. We provide five empirical examples using discrete choice experiments, analysed using conditional logit, heteroscedastic conditional logit, or scale adjusted latent class models. We show that analysts may incorrectly conclude that preferences are consistent across respondents even if they are not, or that classes of respondents may have (in)significant preferences for some or all attributes of the experiment, when they do not. We recommend that future studies employing scale heterogeneity models explicitly state scale factors for all samples, choice contexts, and/or latent scale classes, and report rescaled preference parameters for each of these groups.

Key words: Discrete choice experiments, heteroscedastic conditional logit models, scale adjusted latent class models, interpretation of preferences, best-practice reporting

JEL classifications: C10, C18, C51, Q51

Discrete choice models: scale heterogeneity and why it matters

Katrina J. Davis, Michael Burton and Marit E. Kragt

Introduction

In the discrete choice experiment (Carson and Louviere 2011) literature it is increasingly common to use models which account for heterogeneity in error variance. However, there is a potential issue in the way that results from these models are reported and interpreted. This paper empirically identifies this problem and makes recommendations for best-practice reporting of models which allow for heterogeneity in error variance. Discrete choice experiments are used to investigate preferences. They are particularly useful when eliciting preferences for non-market goods, prospective market goods, or policies which have not yet been implemented. In a discrete choice experiment an individual is asked to choose between different alternatives which are described by attributes (Adamowicz et al. 1998; Swait and Louviere 1993). Choices are then analysed to deconstruct respondents' preferences based on the attributes of the alternatives that they have chosen.

An identified problem with discrete choice experiments is the need to account for heterogeneity when analysing data (Louviere and Eagle 2006). There are two main sources of heterogeneity in discrete choice data. The first is in preferences: groups of respondents like or dislike different alternatives in a systematic and quantifiable way. This systematic component of an individual's utility may be observed through the characteristics of the alternatives and of the individual. Discrete choice experiments aim to measure this preference heterogeneity so that individuals' preferences for policy options or consumer goods can be interpreted in a meaningful way. There has been substantial research effort devoted to estimating heterogeneity in preferences (for example, Birol et al. 2006; Kragt and Bennett 2011; McFadden and Train 2000). The second source of heterogeneity in discrete choice data is due

to heterogeneity in the variance of the error process. Error variance is thought to vary systematically in response to task complexity and the number of choice alternatives or attribute differences (Hensher et al. 1999; Louviere and Eagle 2006).

To identify heterogeneity in preferences we can use random parameter or latent class models. In the former, taste variation amongst individuals is explicitly treated by allowing explanatory variables to vary over individuals (Carlsson et al. 2003; Train 1998). The latter identifies whether groups or classes of respondents share similar preferences (Burton and Rigby 2009). However, both of these models (and all other statistical models in which the dependent variable is latent) will confound estimates of model parameters with error variance (Louviere and Eagle 2006), and hence confound heterogeneity in preferences with heterogeneity in error variance. If error variance is not constant across individuals, then marginal utility estimates which do not account for this variation will be incorrect (Louviere et al. 2002). This implies a need for models which can separate heterogeneity in preference from heterogeneity in error variance so that accurate conclusions about people's preferences can be drawn.

In response to this need, a number of models that can accommodate heterogeneity in error variance have been developed. If two groups of respondents have the same underlying preference parameters, but differ in error variance, heteroscedastic conditional logit (HCL) models can be estimated (Hensher et al. 1999; Hole 2006). Sources of error heterogeneity in these models can include individual or choice task characteristics (Swait and Adamowicz 2001). Other ways of accounting for heterogeneity in error variance include the use of the generalized multinomial logit (G-MNL) model (Fiebig et al. 2010); models for single persons (Louviere et al. 2008); and models that decompose random components (Burke and Reitzig 2007). However, in the context of G-MNL, Hess and Rose (2012) warn that even if heterogeneity in both scale and preferences is explicitly modelled, identification may still be

an issue. This means that what may be attributed to heterogeneity in scale may be indistinguishable from heterogeneity in preferences.

Other models that can account for heterogeneity in error variance are 'scale extended' or 'scale adjusted' latent class (SALC) models (Magidson and Vermunt 2007). These models remain a relatively new area of research but feature in an increasing number of studies including Burke et al. (2010); Rigby et al. (2015); and Thiene et al. (2014). SALC models effectively separate heterogeneity in preferences from heterogeneity in error variance. They do this by identifying latent classes of people who differ in their preferences, as well as latent *scale* classes – groups of people who differ by how systematic (or erratic) they are in their choices. This difference is described by a scale factor which identifies the amount by which the parameter estimates of one group must be rescaled to arrive at the preference parameters appropriate to another group (Magidson and Vermunt 2007).

Scale heterogeneity causes similar issues in best-worst scaling (BWS), a discrete choice experiment which has been receiving increasing research attention (Flynn and Marley 2014). There are three types of BWS, defined as case 1, 2 or 3 (Flynn 2010). A case 1 BWS, also known as the 'object case,' contains a number of choice sets which require the respondent to choose the "best" and "worst" objects or alternatives from varying sets of three or more alternatives (Finn and Louviere 1992). Analysis of the choice data allows each alternative to be rank-ordered on a common scale and assessed on the basis of its relative importance (Marti 2012). In a case 2 BWS, or 'profile case,' respondents are asked to assess the attribute levels of a given profile. Case 3 BWS is the closest to a traditional discrete choice experiment; respondents select their most and least preferred profile or choice set in which levels of attributes are varied (Flynn 2010; Flynn and Marley 2014). Heterogeneity between best and worst choices has previously been modelled using the heteroscedastic SALC model,

which allows variation in both the error variance of different scale classes and between best and worst choices (Rigby et al. 2015).

Models which identify and allow for heterogeneity in error variance, such as HCL or SALC models, address a critical problem in the analysis of discrete choice data – removing the confound between preference and scale heterogeneity. If error variance is heterogeneous across groups of respondents then models which do not account for this heterogeneity will distort utility estimates (Louviere et al. 2002). Scale heterogeneity models therefore offer significant advances in the analysis of discrete choice data. However, potential problems remain with the interpretation of results from these and other such models which need to be fully addressed for the field to advance. This problem is principally linked to the impact that scale (heterogeneity in error variance) has on the significance of preference parameters, and how that will change the interpretation of individuals' preferences. Use of HCL or SALC models could lead the researcher to conclude that preferences between two or more groups of respondents are equivalent if the scale factor is allowed to vary between groups. However, interpretation of preferences may change depending on the scale of the group which is being reported. Once rescaled, the preferences of certain groups may vary substantially; this can lead to misleading conclusions being drawn regarding the significance of whole sample preferences. Depending on the impact of the scale factor, the researcher could mistakenly conclude that the preferences of one group are significant when they are not, or insignificant when they are. In this paper, we contribute to the limited literature on interpreting scale parameters to understand heterogeneity in error variance in discrete choice experiments. Our findings have important implications for the reporting and interpretation of discrete choice experiments when error variance is not uniform over individuals' choices, or across best and worst choices.

In this paper we provide five empirical examples which illustrate how preference data and scale factors can be misinterpreted. All five are discrete choice experiments with two using BWS data. In the next section we provide the theoretical framework for the analysis of discrete choice experiments and the theory behind estimating HCL models. Then we describe the theory behind analysing BWS data with HCL models and heteroscedastic SALC models. In Section 4, we provide a brief description of the methods employed for the analysis of each empirical example before describing each example in detail along with the results from its analysis. Finally, we provide a discussion of the implications of our findings with particular reference to how scale heterogeneity models have been reported in the literature. We conclude with recommendations for best practice reporting of results from these models.

Theoretical framework

Discrete choice experiments

Discrete choice experiments use Random Utility Theory to draw conclusions regarding people's preferences (McFadden 1974). We define latent utility (U) as a function of the vector of attributes (X) and parameters (β) of an alternative m. This expression has a deterministic component V_{im} and a stochastic element captured by the error term ε_{im}, which varies across individuals (i) and alternatives (m). This error term captures all the factors which affect utility and are not captured in the deterministic component, and is assumed to be independently and identically distributed (IID) (Train 2009).

$$U_{im} = V_{im} + \varepsilon_{im} \quad (1)$$

It is well known that there is an identification problem within choice models, such that the underlying preference parameters are conflated with the variance (σ_ε^2) of the error term

(Hensher et al. 1999; Magidson and Vermunt 2007). What are estimated are 'scaled' preference parameters, where, for the conditional logit model, the scale term λ is defined as $\lambda = \frac{\pi}{6^{0.5}\sigma_\varepsilon}$ (Louviere and Eagle 2006). We follow the convention of assuming that λ is normalised to unity in estimation, but strictly, one is identifying only the composite of scale and preference parameters (Hess and Rose 2012). If the error variance (and hence scale term) are not constant across individuals or choices then one runs the risk of confounding heterogeneity in preferences with heterogeneity in error variance (Louviere and Eagle 2006).

If one expects differences in error variance across individuals one can model scale explicitly as a function of observable characteristics:

$$\sigma_{\varepsilon_{im}} \propto \exp(-w_i'\gamma) = \lambda_i \tag{2}$$

Where $exp(w_i'\gamma)$ is the scale factor (λ_i) which is inversely proportional to the standard deviation of the errors, and $w_i{}'\gamma$ is a vector of individual specific characteristics and associated parameters (Rigby et al. 2015).

The probability that alternative *m* is selected from a set of *R* alternatives in the choice set is:

$$\varpi_{im} = \frac{exp[\lambda_i V_{im}]}{\sum_{r=1}^{R} exp[\lambda_i V_{ir}]} \qquad r = 1, m, \ldots, R \tag{3}$$

Equations (1)-(3) describe a model that can allow for differences in error variance over respondents. This model is a heteroscedastic conditional logit model and it assumes constancy in preferences across individuals (Hole 2006).

Best-worst scaling

Analysis of best and worst choices is similarly based on Random Utility Theory. Following Rigby et al. (2015) we define the latent utility (y_{ism}^*), associated with each individual (*i*) and each alternative (*m*), as having a deterministic component $\beta_{m,}$ and a stochastic element

captured by the error term ε_{ism}. Once again the error term is assumed to be IID. We include subscript *s,* which indicates whether the latent utility is derived in the context of a choice which is best or worst, where $s = 1$ if the choice is best, and $s = -1$ if it is worst.

$$y^*_{ism} = \beta_m + \varepsilon_{ims} \qquad (4)$$

Heterogeneity is allowed for in the standard deviation of the error component based on observable characteristics, and best and worst choices.

$$\sigma_{\varepsilon_{ism}} \propto \exp(-w'_{is}\gamma) = \lambda_i \qquad (5)$$

Where $exp(w'_{is}\gamma)$ is the scale factor which is inversely proportional to the standard deviation of the errors, and $w_i{}'\gamma$ is a vector of individual specific characteristics (Vermunt 2013).

We assume sequential best worst ranking: respondents first select the best alternative and then select the worst alternative from the remaining ($M - 1$) alternatives (Vermunt 2013). Thus the probability that alternative m_1 is selected as best ($s = 1$) from a set of R alternatives in the choice set is:

$$\varpi_{ism_1} = \frac{exp[\lambda_i \beta_{m_1}]}{\sum_{r=1}^{R} exp[\lambda_i \beta_r]} \qquad r = 1, m_1, \dots, R;\ s = 1 \qquad (6)$$

The probability that alternative m_2 is selected as worst ($s = -1$) from the $m - 1$ remaining alternatives, conditional upon the choice of best is:

$$\varpi_{ism_2|m_1} = \frac{exp[-\lambda_i \beta_{m_2}]}{\sum_{r \neq m_1}^{R} exp[-\lambda_i \beta_r]} \quad s = -1 \qquad (7)$$

In which the sign of the deterministic component is scaled by -1 as it is the least preferred (worst) preference being chosen.

The probability of selecting m_1 as the best and m_2 as the worst is given by:

$$\varpi_{i,m_1,m_2} = \varpi_{i,m_1} \varpi_{i,m_2|m_1} \qquad (8)$$

Equations (4)-(8) describe a HCL model that can allow for differences in error variance over respondents, and differences in error variance between best and worst choices. Once again this model assumes constancy in preferences; both across individuals and best and worst choices.

This model can be extended to a heteroscedastic SALC model in which both latent scale classes and preference classes can be accommodated. Given *C* latent preference classes (indexed over *j*) and *D* latent scale classes for error variance (indexed over *l*), the probability of selecting best and worst becomes:

$$\varpi_{ism_1|cd} = \frac{exp[\lambda_d \beta_{m_1 c}]}{\sum_{r=1}^{R} exp[\lambda_d \beta_{rc}]} \qquad s = 1 \tag{9}$$

And

$$\varpi_{ism_2|m_1 cd} = \frac{exp[-\lambda_d \beta_{m_{-1} c}]}{\sum_{r \neq m_1}^{R} exp[-\lambda_d \beta_{rc}]} \qquad s = -1 \tag{10}$$

The probability of selecting m_1 as the best and m_2 as the worst is given by:

$$\sum_{cd} P_{cd} \varpi_{i,m_1,m_2|cd} = \sum_{cd} P_{cd}\, \varpi_{i,m_1|cd} \varpi_{i,m_2|m_1,cd} \tag{11}$$

Where P_{cd} is the probability that an individual is a member of preference class *c* and scale class *d*. Membership of scale and preference classes is modelled probabilistically as a function of individual specific characteristics using a multinomial logit functional form:

$$P_{ic} = \frac{exp[z_i' \phi_c]}{\sum_{j=1}^{C} exp[z_i' \phi_j]} \tag{12}$$

$$P_{id} = \frac{exp[z_i' \varphi_d]}{\sum_{l=1}^{D} exp[z_i' \varphi_l]} \tag{13}$$

Where z_i is a vector of individual specific characteristics.

Methods

We present results from a number of empirical examples to illustrate the importance of communicating and accounting for the effect of heterogeneity in the scale factor when interpreting discrete choice data. The examples for which we estimate conditional logit models involve four common steps:

Step 1: Estimation of separate models for two samples of data.

Step 2: Estimation of a restricted model, combining both samples, imposing common preference parameters and error variance, and testing if this restricted model is accepted compared to the models estimated in Step 1.

Step 3: Estimation of a restricted model, combining both samples, imposing common preference parameters but allowing the error variance to differ (i.e. a heteroscedastic conditional logit model) and testing if this restricted model is accepted compared to the models estimated in Step 1.

Step 4: Comparing the implications for interpretation of preferences if different groups are assumed to have a scale factor of 1 (the 'baseline' group) in Step 3.

To illustrate the issue we first conduct a Monte Carlo simulation to show that one could erroneously conclude that two samples have similar preferences as long as they are allowed to differ by a scale factor (for brevity, details are not reported here but are available from the authors). We generate 5000 replications of simulated data, where one sample is assumed to have a linear compensatory utility function and the other sample makes choices at random. For each replication we test a series of null hypotheses regarding the randomness of responses for each sample and whether both samples have the same preference parameters when scale is allowed to vary or not. The number of times that each null hypothesis would be

rejected ($p<0.05$) across the 5000 replications is reported. These results confirm the basic argument: preferences can be misinterpreted if scale heterogeneity is not accounted for.

The first empirical example uses data from a discrete choice experiment where the problem of a subsample with fully random preferences is present. Fully random behaviour may be rare in real applications, therefore the second empirical example uses a simulated data set where in one subsample only a subset of parameters are truly significant, but one may conclude that both subsamples have equivalent preferences.

The third empirical example is from a real case 1 BWS study where models of best and worst choices imply that respondents have the same preference parameters once one allows for scale heterogeneity. In this example we follow the same analysis outlined above, but instead of comparing two subsamples of different individuals, we compare best and worst choices for the same individuals. In the fourth empirical example we extend the analysis undertaken in example 3 by conducting a SALC analysis on the same data set to evaluate whether there are latent classes of respondents with different error variances. Preference parameters are rescaled and reported for both best and worst choice models and for all scale classes.

The fifth empirical example explores the circumstances under which misinterpretation of preferences is likely to occur. This example identifies that it is the precision with which one estimates scale factors, and not the size of the scale factor, which has the greatest impact on the interpretation of preference.

All conditional logit and heteroscedastic conditional logit models were analysed with Stata 13 (StataCorp 2013). SALC analyses were performed using Latent GOLD Choice 5.0 (Vermunt and Magidson 2005, 2014). Data sets and all model code are available from the authors.

Results: Empirical examples

Empirical example 1: interpretation problems in a real data set

Our first data set involves a discrete choice experiment which explores the preferences of a sample of 318 people regarding a marine biodiversity offset package with multiple attributes. This was a forced choice: the framing was that a project would take place. Respondents were asked to choose the offset package that they would prefer to see implemented to achieve no-net ecological loss. Full details of this survey are reported in Rogers et al. (2014).

Respondents had three alternatives in each choice set (for an example choice set see Appendix A). Attributes were the location of the offset, which was either in Western Australia (the baseline), Queensland (AU), New Zealand or China; the species of bird being protected, which was either the *Ruddy Turnstone* (baseline) or *Eastern Curlew*; and the proportion of offset that was 'direct'. For the purpose of the current analysis, individuals within the sample are split into two sub-samples A and B (n=250 and 68 respectively). Note that this assignment to sub-samples was not part of the original data collection process and has been undertaken purely for the purpose of this exposition: individuals who appear to have made their choices at random have been identified *ex-post* and the sample split accordingly. The question for the current analysis is whether one would classify the two samples as having the same preference structure.

Table I reports the results for separate conditional logit models for Samples A and B, a stacked data set that ignores potential heterogeneity in scale, and two heteroscedastic conditional logit models that differ only in which subsample is used as the baseline (scale normalised to unity). Note that here, and in all subsequent results, we report the scale factor (λ, see Theoretical framework) as opposed to the underlying parameter estimated by the software.

Table I. Conditional logistic and heteroscedastic conditional logistic regression results for a discrete choice experiment eliciting people's preferences for a marine offset package.

Attributes[a]	Sample A	Sample B	Stacked A and B	Heteroscedastic: Sample A[b]	Heteroscedastic: Sample B[c]
Queensland	-2.51 ***	0.35 *	-1.70 ***	-2.51 ***	-3.45E-05
New Zealand	-0.90 ***	-0.04	-0.72 ***	-0.90 ***	-1.24E-05
China	-1.65 ***	0.27	-1.13 ***	-1.65 ***	-2.27E-05
E. Curlew	-0.18 ***	0.05	-0.13 ***	-0.18 ***	-2.41E-06
Percent of direct offset	0.04	-0.33	-0.08	0.04	5.78E-07
Scale factors					
Sample A				1	72671.17 **
Sample B				2.34E-08[d]	1
Individuals	250	68	318	318	318
Observations	4452	1182	5634	5634	5634
Log likelihood	-1305.79	-428.68	-1838.98	-1738.65	-1738.65
LR chi^2	649.09	8.34	448.43	200.67	200.66
Prob > chi^2	0.00	0.14	0.00	0.00	0.00

*** P>|z| <0.01
[a]Location baseline is Western Australia, and the species' baseline is to protect the *Ruddy Turnstone*.
[b]Normalised to Sample A, values for Sample B can be calculated by multiplying preference parameters by 2.34E-08.
[c]Normalised to Sample B, values for Sample A can be calculated by multiplying preference parameters by 72671.17.
[d]In this case the heteroscedastic conditional logit model can assign any sufficiently small (Sample A) or large (Sample B) scale factor to arrive at the alternative preference coefficients, hence there is very poor precision in the estimation of these scale factors.

The results of a likelihood ratio test comparing the fit of stacked data for samples A and B, with individually estimated models of Samples A and B, roundly reject the null hypothesis that sample A and B can be restricted to have equal parameters ($\chi_{df=4} = 209.00$; $\chi_{0.05,\ df=4} = 11.07$; p-value < 0.001). The results of a second likelihood ratio test, which compared the fit of the stacked data for samples A and B while allowing the scale parameter to vary, show that we cannot reject the restriction that the preference parameters are equal ($\chi_{df=4} = 8.34$; $\chi_{0.05,\ df=4} = 9.49$; p-value = 0.08).

It is true that the parameter estimates for sample B are consistent with scaled values of sample A, but only if the scaling is so extreme that the parameters become insignificantly different from zero. Although this data comes from a real data set, it is unlikely that one would conclude that preferences were equivalent if undertaking this analysis: Sample B

estimates are clearly not significant, and the scale factor is a very large negative value implying a dramatic rescaling (2.34E-08).

Empirical example 2: Simulated dataset where the researcher could erroneously accept that samples had equivalent preferences

Purely random responses like those seen in empirical example 1 may be rare. However, there may be data which contains subsamples who are clearly not random in behaviour, and have different preferences – but who might be misconstrued as having identical preferences due to the variance heterogeneity confound. In this example we use another simulated dataset to explore interpretation problems which can arise when comparing two samples with different utility functions and one has more systematic preferences. We created two samples: A and B, each with 5000 individuals. Both samples chose between three alternatives, each described by four attributes: X_1, X_2, X_3 and X_4. Sample A was the more systematic group with the following latent utility function:

$$U_A = 0.5X_1 - 0.5X_2 + 2X_3 - 2X_4 - 4.5\varepsilon \qquad (14)$$

Sample B was the less systematic group with the following latent utility function:

$$U_B = 0X_1 + 0X_2 + X_3 - X_4 - 8\varepsilon \qquad (15)$$

Estimation of conditional logit models for each sample individually and for stacked data for both samples are shown in Table II. The results for a heteroscedastic conditional logit model where stacked data for Samples A and B were allowed to vary by a scale factor are also presented.

Table II. Conditional logistic and heteroscedastic conditional logistic regression results for a simulated data set with two samples: A and B.

Attributes	Sample A	Sample B	Stacked A and B	Heteroscedastic: Sample A[a]	Heteroscedastic: Sample B[b]
X_1	0.12 **	0.03	0.08 *	0.12 **	0.04
X_2	-0.12 **	-0.04	-0.08 *	-0.12 **	-0.04
X_3	0.43 ***	0.14 **	0.28 ***	0.43 ***	0.14 ***
X_4	-0.41 ***	-0.13 **	-0.27 ***	-0.41 ***	-0.13 ***
Scale factors					
Sample A				1	3.16 ***
Sample B				0.32 ***	1
Individuals	5000	5000	10000	10000	10000
Observations	15000	15000	30000	30000	30000
Log likelihood	-5441.26	-5487.79	-10941.14	-10929.05	-10929.05
LR chi^2	103.61	10.54	89.97	24.17	24.17
Prob > chi^2	0.00	0.03	0.00	0.00	0.00

*** P>|z| <0.01

[a]Normalised to Sample A, values for Sample B can be calculated by multiplying preference parameters by 0.32.

[b]Normalised to Sample B, values for Sample A can be calculated by multiplying preference parameters by 3.16.

The results of a likelihood ratio test comparing the fit of stacked data for Samples A and B, with Samples A and B when estimated individually roundly reject the null hypothesis that sample A and B can be restricted to have equal parameters ($\chi_{df=4} = 24.19$; $\chi_{0.05,\, df=4} = 9.48$; p-value = 0.0001). The results of a second likelihood ratio test, which compared the fit of the stacked data for samples A and B while allowing the scale parameter to vary, show that we cannot reject the restriction that the preference parameters are equal ($\chi_{df=3} = 0.02$; $\chi_{0.05,\, 3} = 7.81$; p-value = 0.9995).

What one interprets as significant depends on the normalisation of the scale factor. Where this is normalised to Sample A (Heteroscedastic: Sample A) the conclusion would be that all preference parameters are significant, but that Sample B is less consistent in their choices ($\lambda = 0.32$). However, if preference parameters are rescaled to Sample B (Heteroscedastic: Sample B), then it becomes apparent that respondents in this sample (B) are not just less consistent in their choices, but actually place a zero value upon attributes X_1 and X_2.

Correspondingly, this empirical example demonstrates not only that Sample A's preferences are not equivalent to Sample B's (despite the apparent ability to stack data sets and model them as if they were), but also that the interpretation of the data (statistical significance of preference parameters) will depend on whether the model is normalised (and reported) to Sample A or Sample B.

Empirical example 3: A best-worst scaling data set where analysis would suggest that best and worst choices were the same

This example uses data from a case 1 BWS survey completed by members of artisanal fisher organisations in the central marine region of Chile (Davis et al. 2015). The survey asked respondents what they thought was the most important reason not to monitor marine management areas. Management areas are geographically specific coastal marine areas where territorial user rights for fisheries (TURFs) have been assigned to artisanal fisher organisations by the Chilean government for extraction of benthic (bottom-dwelling) resources.

The survey was completed by 52 respondents who were a mixture of management and non-management members of their fisher organisation. The survey was constructed using a balanced incomplete block design (Cochran and Cox 1950). There were seven alternative reasons not to enforce, four alternatives shown in each BWS question, and each alternative was shown four times throughout the survey (an example BWS question is shown in Appendix A).

Estimation of conditional logit models for best and worst choices and for stacked best and worst choice data are shown in Table III. The results for a heteroscedastic conditional logit

model where stacked data for best and worst choices were allowed to vary by a scale factor are also presented.

Table III. Conditional logistic and heteroscedastic conditional logistic regression results for best and worst choice data from a survey investigating what artisanal fishers in Chile thought were the most and least important reasons not to monitor marine management areas.

Reasons not to monitor[a]	Best choices	Worst choices	Stacked best & worst choices	Heteroscedastic: Best choices[b]	Heteroscedastic: Worst choices[c]
1	0.11	0.12	0.11	0.13	0.05
2	0.30	0.26	0.28 *	0.33 *	0.12
3	-0.15	-0.10	-0.14	-0.17	-0.06
4	-0.16	0.06	-0.04	-0.11	-0.04
5	0.69 ***	-0.16	0.43 ***	0.62 ***	0.23 ***
7	-1.46 ***	-0.66 ***	-0.83 ***	-1.56 ***	-0.58 ***
Scale Factors					
Best choices				1	2.67 ***
Worst choices				0.37 ***	1
Observations	364	273	637	637	637
Log likelihood	-465.17	-386.88	-863.46	-855.25	-855.25
LR chi^2	78.88	26.03	82.09	16.43	16.43
Prob > chi^2	0.000	0.000	0.000	0.000	0.000

*** P>|z| <0.01
[a]Reason 6 was the baseline.
[b]Normalised to best choices, values for worst choices are calculated by rescaling parameter coefficients by a factor of 0.37.
[c]Normalised to worst choices, values for best choice are calculated by rescaling parameter coefficients by 2.67.

A likelihood ratio test on the stacked best and worst choice model rejected the null hypothesis that the coefficients of models of best and worst choices were the same ($\chi_{df=6} = 22.82$; $\chi_{0.05,\,6} = 12.59$; p-value = 0.0009). This implies that marginal utilities differ between best and worst choices.

The heteroscedastic conditional logit model of best and worst choices imposes that marginal utilities for both best and worst choices are the same, but allows the error variance between best and worst choices to be different. A likelihood ratio test for restricting data did not reject the null hypothesis that the coefficients of models of best and worst choices were the same

across both decisions (i.e. the model could predict both best and worst choices) ($\chi_{df=5} = 6.39$; $\chi_{0.05,5} = 11.07$; p-value = 0.27).

The results from Table III would suggest that best and worst choices can be modelled together provided that an adjustment in scale is made for worst choices. Rescaling the preference parameters of best choices by a scale factor (0.37) returns preference parameters for worst choices (Heteroscedastic: Worst choices); these preference parameters are smaller than for best choices, but still significant. We could therefore conclude that preferences parameters for both best and worst choices are roughly equivalent, but that respondents were less consistent when making their worst choices. Although analysis in this example has tested for differences in error variance between best and worst choices, it has not explored whether groups of respondents were more or less consistent in their choices overall.

Empirical example 4: Preference parameters rescaled to best and worst choices and different scale classes may require different interpretations

We also analysed the data set from example 3 using Latent GOLD Choice 5.0 (Vermunt and Magidson 2005, 2014) to assess whether latent scale classes were present in the data[1] – groups of respondents who were or more less similar in their error variance. This analysis allowed us to simultaneously model two sources of heterogeneity in error variance: between best and worst choices (which was deterministic) and generic choice consistency (which was modelled as membership of a latent scale class).

Results suggested that there were two latent scale classes present (based on CAIC). We correspondingly report four different representations of results for the same model (Table IV). All models are statistically identical (as indicated by the LL values); the only difference

[1]We also tested for differences in preference classes, but these not could be robustly identified.

is the normalisation of the error variance for best or worst choices and for different latent scale classes. Scale class 1 account for 25% of the sample, and have a larger scale/smaller variance, i.e. are more consistent in their choices.

Table IV. Heteroscedastic scale adjusted latent class model of case 1 best and worst choice data from a survey investigating what artisanal fishers in Chile thought was the most and least important reason not to monitor marine management areas.

Reasons not to monitor[a]	4a	4b	4c	4d
5	1.296 ***	0.084	0.001	0.022
4	0.986 **	0.064	0.001	0.017
2	0.607	0.039	0.001	0.011
3	-0.046	-0.003	0.000	-0.001
1	-0.262	-0.017	0.000	-0.005
7	-73.035 ***[b]	-4.712 ***	-0.082	-1.265 ***
Scale factors				
Best, Scale class 1	1	15.501 ***	895.068 ***	57.743 ***
Worst, Scale class 1	0.065 ***	1	57.743 ***	3.725 ***
Worst, Scale class 2	0.001 ***	0.017 ***	1	0.065 ***
Best, Scale class 2	0.017 ***	0.268 ***	15.501 ***	1
Class Membership				
Scale class 1 (25%)	0.000	0.000	-1.054 ***	-1.054 ***
Scale class 2 (75%)	1.054 ***	1.054 ***	0.00	0.000
Observations	637	637	637	637
Log likelihood	-855.250	-855.250	-855.250	-855.250
LR chi^2	16.430	16.430	16.430	16.430
Prob > chi^2	0.0001	0.0001	0.0001	0.0001

*** P>|z| <0.01

[a]Reason 6 was the baseline.

[b]Not well identified by Latent GOLD choice 5.0 (Vermunt and Magidson 2005, 2014) – any sufficiently large negative coefficient will ensure that the predicted probability approximates zero; there is consequently very poor precision in the estimation of this coefficient.

In Model 4a (Table IV), the choice and scale class normalised to have a scale factor of 1 are the "best" choices and latent scale class 1. We note a strong aversion to alternative 7, while alternatives 5 and 4 seem to be ranked more highly than 6 (the baseline). Alternatives 2, 3 and 1 are ranked similarly to 6 (not significantly different from zero).

Model 4b (Table IV) is exactly the same statistical model, but normalised so the scale factor of 1 relates to scale class 1 and the "worst" choices. All values in Model 4b can be retrieved from Model 4a by rescaling by a factor of 0.065. As one would expect, this scale factor strongly influences the significance of preference parameters; in Model 4a we know that the error variance is much higher for the "worst" answers, and hence the rescaled parameters (Model 4b) are closer to and may not be significantly different to zero. Consequently, in Model 4b the utility function that underlies the "worst" answer responses looks much more like a random process, apart from the strong aversion to alternative 7 which is still significant in the "worst" choices as it was in the "best".

Model 4c (Table IV) is still normalised so that worst choices have a scale of 1, but now shifts the scale class normalisation from scale class 1 to scale class 2, who represent 75% of the sample. For this category none of the parameters are significantly different from zero, implying that choices are random.

Model 4d (Table IV) is the final analysis in the set: scale is normalised to equal 1 for the "best" choices, and for scale class 2. Here we see no discrimination between alternatives 1-6, i.e. respondents are assumed to be indifferent between these alternatives. Once again alternative 7 is significantly not preferred.

Model 4a (parameter estimates normalised to best choices and for scale class 1) presented a relatively positive picture: the model seemed to be detecting differences in preference ranking, as implied by the significant differences in marginal utility parameters for different reasons not to monitor, and there were some interesting effects manifesting themselves in terms of error variances. However, a closer inspection of the data suggested that alternative interpretations were hidden within this result: for 75% of the sample (scale class 2) there was no useful information revealed in their worst choices, and only an aversion to alternative 7 was significant in their best choices. For the remaining 25% only an aversion to alternative 7

was significant in their worst choices and any further differentiation in marginal utility parameters was limited to best choices. The implications of these results are that only a small proportion (25%) of the population was making systematic choices between different reasons not to enforce. However, unless preference parameters were rescaled and reported for both best and worst choices and for all latent scale classes, this conclusion may have been overlooked.

Empirical example 5: Under what circumstances may this be an issue?

In the previous examples we have shown that there is a potential issue with the reporting of models that involve heterogeneity in scale: the rescaling implied by the scale parameter may not always be inconsequential in the interpretation of the behavioural implications of the model for some segments of the population. A relevant question is: when is this likely to be an issue? The variable of concern is the degree of precision of the estimated scale parameter. Note from equation 3 that the scale parameter λ enters the probability model multiplicatively with the preference parameters. Conventionally, one scale class has the scale parameter fixed at 1 for identification, and again, by convention, the preference parameters are reported for that class. The preference parameters for the other class(es) can be retrieved by multiplying by the estimated scale parameter(s). Given that estimates of both preference and scale parameters are random variables, the distribution of the scaled preference parameters will be given by the distribution of the product of two random distributions. Given the requirement that the scale parameter is positive, estimation usually specifies equation (2), so that the distribution of the scaled preference parameter will be the product of a normal and log-normal distribution.

There is no analytical solution to the moments of this distribution, but some general observations can be made. If the scale parameter is identified with absolute certainty, then the

significance of the preference parameters for all scale classes will be identical: the application of the scaling will alter both the mean and standard error of the parameters equally, leaving the z-statistic unchanged. Thus even if the magnitude of those coefficients is changed, the behavioural interpretation of which attributes are significant will not be. If one applies the convention that the class with the most consistent preferences has a scale normalised to 1 (so all other estimated scale values are less than 1), as the standard error of the scale parameter increases (i.e. the precision of the estimate of the scale parameter falls) the precision of the scaled preferences also falls.

The implications of these observations are demonstrated below in Table V and Table VI, which report results for two (simulated) data sets with a single preference class, but two latent scale classes (code available from the authors). For data set 1 (Table V) the scale parameter is relatively precisely estimated (z-statistic = -4.07) and there is little change in the implied significance of the parameters between Models 1 and 2 despite the rescaling effect. This is not the case for data set 2 (Table VI) which has a similar scale parameter (-1.56 in Model 2 compared to -1.52 in Model 1), but which is less precisely estimated (z-statistic = -3.32). The impact of this uncertainty in the estimation of the scale parameter is seen in the difference between the significance of Model 3 and Model 4 preference parameters (Table VI). This example highlights where issues of interpretation may arise: when the scale parameter is relatively imprecisely identified. Table VI emphasises our main conclusion: reporting of Model 3 might lead one to conclude "all attributes are significant, and scale class 2 exhibits higher scale heterogeneity" when a more accurate evaluation would be "for members of scale class 1, all attributes are significant, while for scale class 2 the scale heterogeneity is sufficiently large to make choices indistinguishable from random behaviour".

Table V. Results for (simulated) data set 1: one preference class, but two latent scale classes.

	Model 1		Model 2	
Attributes	Coeff.	z-statistic	Coeff.	z-statistic
X_1	-0.66	-5.17	-0.15	-2.12
X_2	0.60	4.48	0.13	2.03
X_3	-1.98	-7.98	-0.43	-2.21
X_4	-1.82	-7.56	-0.40	-2.19
Scale parameters				
Scale class 1	0		1.52	4.07
Scale class 2	-1.52	-4.07	0	
Log likelihood	-4092.49		-4092.49	

Table VI. Results for (simulated) data set 2: one preference class, but two latent scale classes.

	Model 3		Model 4	
Attributes	Coeff.	z-statistic	Coeff.	z-statistic
X_1	-0.65	-5.17	-0.14	-1.76
X_2	0.56	4.39	0.12	1.70
X_3	-1.88	-7.73	-0.40	-1.80
X_4	-1.73	-7.33	-0.37	-1.79
Scale parameters				
Scale class 1	0		1.56	3.32
Scale class 2	-1.56	-3.32	0	
Log likelihood	-4098.49		-4098.49	

The impact of uncertainty in the scale parameter estimate on the significance of a scaled preference parameter estimate is shown in Figure 1. We assume that we have an estimated preference parameter (β) for the scale class with scale normalised to 1, which is normally distributed with a mean of 1 and SE of 0.33 (implying a z-statistic of 3.33). We further assume that the scale factor has been estimated as a log normal function i.e. λ=exp(z), where z is normally distributed, giving an implied mean μ and standard deviation σ of the scale factor. The expected value of the scaled preference parameter is given by $S\beta = \beta \times \lambda$, i.e. the product of a normal and log normal distribution. We evaluate the distribution of $S\beta$ by a simulation of 1000 draws from the two distributions, and evaluating the mean and SE of the resulting distribution, and its implied z value, for all combinations of 10 values for μ and σ,

giving us 100 data points showing the relationship between the mean and z-statistic for the scale factor, and the implied z-statistic for the scaled preference coefficient. This is replicated in a Monte Carlo simulation 1000 times for each data point, and the mean values are reported in Figure 1. When the z-statistic of the scale factor is large, then the estimated z-statistic for the scaled preference parameter asymptotes to a value of 3.33, as expected. As the precision of the scale factor falls, the z-statistic of the scaled preference parameter also falls, confirming the results seen in Table V and Table VI. Changing the mean of the scale factor has no effect on the relationship.

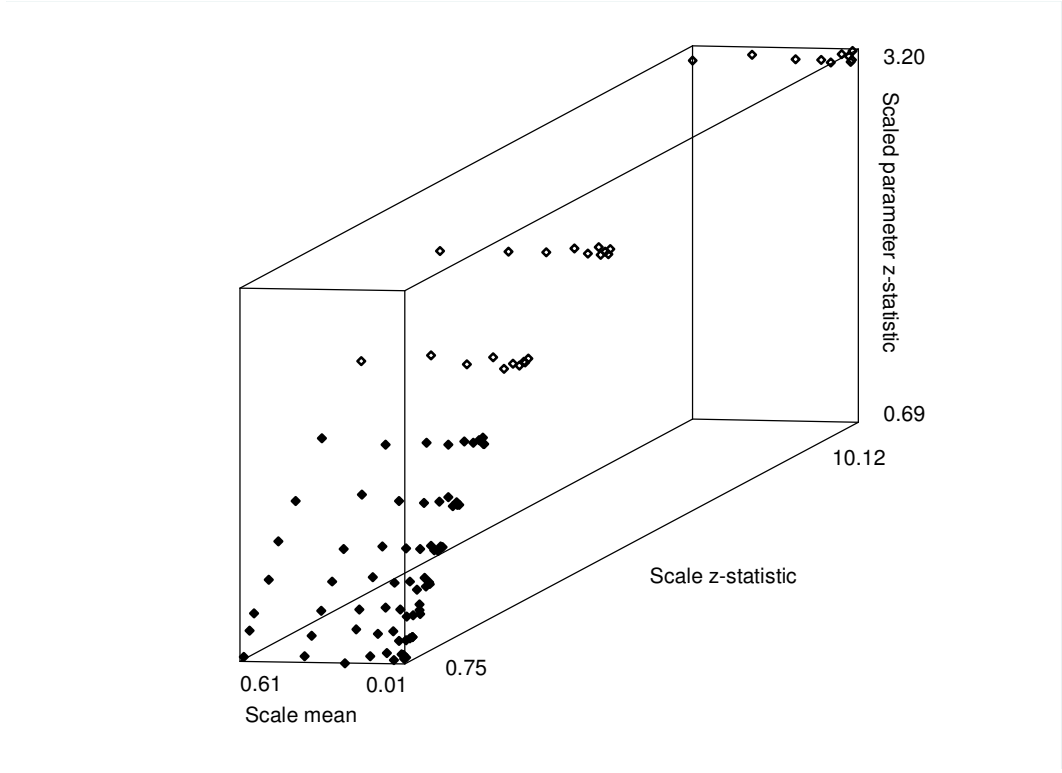

Figure 1. The relationship between the size of the scale factor, standard error of the scale factor, and the z-statistic of an estimated parameter. The symbol ♦ indicates scaled parameter z-statistics <1.96.

Discussion

We have demonstrated that preference parameters from discrete choice data can be misinterpreted if due care is not taken when communicating the impact of scale factors. The importance of the scale parameter has been previously identified (Louviere and Eagle 2006),

but its potential to affect the interpretation of discrete choice data has not been empirically demonstrated before. We show that interpretation can be a concern when using models which allow for scale heterogeneity, such as HCL or SALC models. These models allow preferences of two or more groups to be combined into a single model providing the preference parameters of different groups are allowed to vary by a scale factor. However, the interpretation of preference parameters for this combined model (or their statistical significance) may change depending on which group is selected as the baseline. The impact of this change is likely to depend on how precisely scale and preference parameters are estimated (e.g. z-statistic). We illustrated the potential for preference data to be misinterpreted in five empirical examples which were analysed with HCL and heteroscedastic SALC models. These examples illustrate how it is possible to draw misleading conclusions regarding respondent's preferences if the estimated parameters and their significance are not explicitly considered for all scale classes.

These results have implications for the interpretation of a number of analyses which have now been undertaken using SALC models – both in a standard discrete choice context, for example Mueller Loose et al. (2013), and when using BWS, for example Rigby et al. (2015). Closer examination of how these, and other publications, report and interpret scale factors reinforces our conclusions regarding the potential for misinterpretation and the need for best practice reporting of models when the error variance of different groups is allowed to vary.

Of the 15 papers that we reviewed that use SALC models[2], most report scale factors and scale class membership. However, none extend their discussion of the potential impacts of heterogeneity in error variance across latent scale classes beyond noting that one latent scale class is more or less "consistent" in their choices than the other (see, for example Mueller et

[2]Bartczak and Meyerhoff (2013), Burke, et al. (2010), Campbell, et al. (2011), Flynn, et al. (2013), Flynn, et al. (2010), Glenk, et al. (2012), Islam (2014), Louviere, et al. (2013), Mueller Loose, et al. (2013), Mueller, et al. (2010), Rigby, et al. (2014), Sagebiel and Rommel (2014), Schlereth, et al. (2012), Tapsuwan, et al. (2014), and Thiene, et al. (2012.

al. 2010). This omission of the impacts of scale heterogeneity is also observed in a BWS context. For example, Rigby et al. (2015) conduct a heteroscedastic SALC analysis of BWS choice data regarding researchers' journal preferences. They mention that respondents were less consistent when identifying worst choices compared to best choices, but do not rescale preference parameters for worst choices to ensure that their interpretation of preferences was robust to rescaling.

It is important to note that heterogeneity in error variance across different samples, choices, and/or latent scale classes will not cause interpretation problems in all analyses. For some analyses[3] it is unlikely that the authors' interpretation of results would change if they rescaled preference parameters for the alternate latent scale class. This is likely to be the case where the scale and preference parameters are precisely estimated (higher z-statistics). However, where scale and preference parameters are less significant (lower z-statistics) interpretation is more likely to be sensitive to rescaling. Preference parameters which are significant for one scale class may become insignificant once rescaled for the alternate scale class. Despite this potential for interpretation to change, none of the 15 SALC papers that we reviewed speculate on how their results – or conclusions drawn from those results – might change if they rescaled preference parameters for alternate latent scale classes or "worst" choices. Some of these papers do not report scale factors or parameters despite using a SALC model (e.g. Islam 2014) – for these papers the reader cannot interpret the significance of preference parameters or whether the authors' conclusions would be robust to rescaling.

We have shown potential problems that can arise when interpreting the significance of preference parameters without accounting for scale. If care is not taken to report and interpret the impact of scale factors, erroneous conclusions may be drawn regarding individuals'

[3]Re-estimation and rescaling of preference parameters to alternate groups caused only marginal changes to interpretation of results presented in Rigby et al. (2015), Tapsuwan et al. (2014), and Bartczak and Meyerhoff (2013). The z-statistics for scale factors in these analyses were -11.65, 3.83, and 4.35 respectively.

preferences for different alternatives. This potential for misinterpretation will not exist in equal measure for all analyses, but cannot be ascertained *a priori*. Thus future studies should explicitly state scale factors for all samples, choice contexts, and/or latent scale classes; and report rescaled preference parameters for the same. Otherwise, it is not clear to either authors or readers whether preference data is being accurately interpreted or portrayed.

Appendix A

This section provides the choice sets used in empirical examples 1 and 3.

A.1

Choice set used for the discrete choice experiment in empirical example 2 regarding preferences for a marine biodiversity offset package (Figure A.1). The survey framing was that a project would take place, and respondents were asked only for their preferences regarding the attributes of the offset package to achieve no-net-loss ecologically. An example of the choice question is given below.

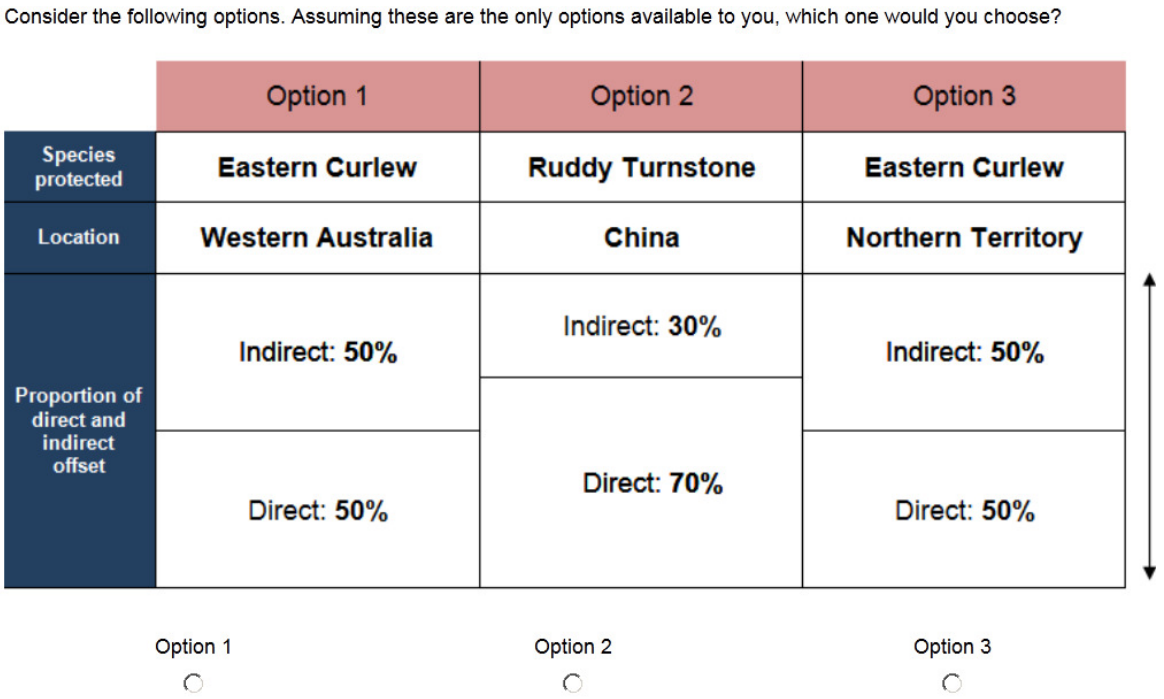

Figure A.1. Example of the offsets discrete choice question from empirical example 2.

A.2

Best-worst scaling (BWS) question used for the discrete choice experiment in empirical example 4 (Figure A.2). The survey asked respondents what they thought was the most and least important reason not to monitor marine management areas, and was completed by members of artisanal fisher organisations in the central marine region of Chile (Davis et al. 2015). Management areas are geographically specific coastal marine areas where territorial user rights for fisheries (TURFs) have been assigned to artisanal fisher organisations by the Chilean government for extraction of benthic (bottom-dwelling) resources.

What **MOST** affects your decision	The decision not to monitor a management area	What **LEAST** affects your decision
✓	**1.** We feel too uncomfortable monitoring or denouncing poachers	
	2. Government punishments for poachers are not effective	
	3. Monitoring represents a high personal risk for guards	
	4. The management area is too far from the *caleta*	✓

Figure A.2. Example of best-worst choice set from empirical example 4. Respondents were asked what they thought was the most and least important reasons not to monitor marine management areas.

References

Adamowicz, W., Boxall, P., Williams, M., Louviere, J., 1998. Stated Preference Approaches for Measuring Passive Use Values: Choice Experiments and Contingent Valuation. American Journal of Agricultural Economics 80, 64-75.

Bartczak, A., Meyerhoff, J., 2013. Valuing the chances of survival of two distinct Eurasian lynx populations in Poland – Do people want to keep the doors open? Journal of Environmental Management 129, 73-80.

Birol, E., Karousakis, K., Koundouri, P., 2006. Using a choice experiment to account for preference heterogeneity in wetland attributes: The case of Cheimaditida wetland in Greece. Ecological Economics 60, 145-156.

Burke, P.F., Burton, C., Huybers, T., Islam, T., Louviere, J.J., Wise, C., 2010. The Scale-Adusted Latent Class Model: Application to Museum Visitation. Tourism Analysis 15, 147-165.

Burke, P.F., Reitzig, M., 2007. Measuring patent assessment quality—Analyzing the degree and kind of (in)consistency in patent offices' decision making. Research Policy 36, 1404-1430.

Burton, M., Rigby, D., 2009. Hurdle and Latent Class Approaches to Serial Non-Participation in Choice Models. Environmental and Resource Economics 42, 211-226.

Campbell, D., Hensher, D.A., Scarpa, R., 2011. Non-attendance to attributes in environmental choice analysis: a latent class specification. Journal of Environmental Planning and Management 54, 1061-1076.

Carlsson, F., Frykblom, P., Liljenstolpe, C., 2003. Valuing wetland attributes: an application of choice experiments. Ecological Economics 47, 95-103.

Carson, R.T., Louviere, J.J., 2011. A Common Nomenclature for Stated Preference Elicitation Approaches. Environmental and Resource Economics 49, 539-559.

Cochran, W.G., Cox, G.M., 1950. Experimental designs (Wiley mathematical statistics series). Wiley, New York.

Davis, K., J, Kragt, M.E., Gelcich, S., Schilizzi, S., Pannell, D.J., 2015. What prevents fishers from enforcing their user rights? Working Paper 1510. School of Agricultrual and Resource Economics, University of Western Australia, Crawley, Australia.

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.

Finn, A., Louviere, J.J., 1992. Determining the appropriate response to evidence of public concern: the case of food safety. Journal of Public Policy & Marketing 11, 12-25.

Flynn, T.N., 2010. Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. Expert Review of Pharmacoeconomics & Outcomes Research 10, 259-267.

Flynn, T.N., Huynh, E., Peters, T.J., Al-Janabi, H., Clemens, S., Moody, A., Coast, J., 2013. Scoring the icecap-a capability instrument. Estimation of a uk general population tariff. Health Economics, 258-269.

Flynn, T.N., Louviere, J.J., Peters, T.J., Coast, J., 2010. Using discrete choice experiments to understand preferences for quality of life. Variance-scale heterogeneity matters. Social Science & Medicine 70, 1957-1965.

Flynn, T.N., Marley, A.A.J., 2014. Best worst scaling: Theory and methods, In Handbook of Choice Modelling. eds S. Hess, A. Daly, pp. 178-201. Edward Elgar Publishing, UK.

Glenk, K., Hall, C., Liebe, U., Meyerhoff, J., 2012. Preferences of Scotch malt whisky consumers for changes in pesticide use and origin of barley. Food Policy 37, 719-731.

Hensher, D., Louviere, J., Swait, J., 1999. Combining sources of preference data. Journal of Econometrics 89, 197-221.
Hess, S., Rose, J., 2012. Can scale and coefficient heterogeneity be separated in random coefficients models? Transportation 39, 1225-1239.
Hole, A.R., 2006. Small-sample properties of tests for heteroscedasticity in the conditional logit model. Economics Bulletin 3, 1-14.
Islam, T., 2014. Household level innovation diffusion model of photo-voltaic (PV) solar cells from stated preference data. Energy Policy 65, 340-350.
Kragt, M.E., Bennett, J.W., 2011. Using choice experiments to value catchment and estuary health in Tasmania with individual preference heterogeneity*. Australian Journal of Agricultural and Resource Economics 55, 159-179.
Louviere, J., Eagle, T., 2006. Confound it! That pesky little scale constant messes up our convenient assumptions, In Proceedings of the Sawtooth Software Conference. pp. 211-228.
Louviere, J., Lings, I., Islam, T., Gudergan, S., Flynn, T., 2013. An introduction to the application of (case 1) best–worst scaling in marketing research. International Journal of Research in Marketing 30, 292-303.
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.
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, 128-164.
Magidson, J., Vermunt, J.K., 2007. Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference, In Sawtooth software conference. p. 139.
Marti, J., 2012. A best–worst scaling survey of adolescents' level of concern for health and non-health consequences of smoking. Social Science & Medicine 75, 87-97.
McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior, In Frontiers in Econometrics. ed. P. Zarembka, pp. 105-142. Academic Press, New York.
McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. Journal of applied Econometrics 15, 447-470.
Mueller Loose, S., Peschel, A., Grebitus, C., 2013. Quantifying effects of convenience and product packaging on consumer preferences and market share of seafood products: The case of oysters. Food Quality and Preference 28, 492-504.
Mueller, S., Lockshin, L., Saltman, Y., Blanford, J., 2010. Message on a bottle: The relative influence of wine back label information on wine choice. Food Quality and Preference 21, 22-32.
Rigby, D., Burton, M., Lusk, J.L., 2015. Journals, Preferences, and Publishing in Agricultural and Environmental Economics. American Journal of Agricultural Economics 97, 490-509.
Rogers, A.A., Burton, M., Richert, C., Kay, A., 2014. Community acceptance of marine biodiversity offsets in Australia: a pilot study. Marine Biodiversity Hub, National Environmental Research Program, Perth.
Sagebiel, J., Rommel, K., 2014. Preferences for electricity supply attributes in emerging megacities — Policy implications from a discrete choice experiment of private households in Hyderabad, India. Energy for Sustainable Development 21, 89-99.
Schlereth, C., Eckert, C., Skiera, B., 2012. Using discrete choice experiments to estimate willingness-to-pay intervals. Marketing Letters 23, 761-776.
StataCorp, 2013. Stata Statistical Software: Release 13. StataCorp LP, College Station.

Swait, J., Adamowicz, W., 2001. 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., 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.
Tapsuwan, S., Burton, M., Mankad, A., Tucker, D., Greenhill, M., 2014. Adapting to Less Water: Household Willingness to Pay for Decentralised Water Systems in Urban Australia. Water Resources Management 28, 1111-1125.
Thiene, M., Meyerhoff, J., De Salvo, M., 2012. Scale and taste heterogeneity for forest biodiversity: Models of serial nonparticipation and their effects. Journal of Forest Economics 18, 355-369.
Thiene, M., Scarpa, R., Louviere, J., 2014. Addressing Preference Heterogeneity, Multiple Scales and Attribute Attendance with a Correlated Finite Mixing Model of Tap Water Choice. Environmental and Resource Economics, 1-20.
Train, K.E., 1998. Recreation Demand Models with Taste Differences over People. Land Economics 74, 230-239.
Train, K.E., 2009. Discrete choice methods with simulation. Cambridge University Press.
Vermunt, J.K., 2013. Categorical response data, In The SAGE Handbook of Multilevel Modeling. eds M.A. Scott, J.S. Simonoff, B.D. Marx, pp. 287-298. Sage, Thousand Oaks, CA.
Vermunt, J.K., Magidson, J., 2005. Latent GOLD Choice 4.0 User's Manual. Statistical Innovations Inc., Belmont Massachusetts.
Vermunt, J.K., Magidson, J., 2014. Upgrade Manual for Latent GOLD Choice 5.0: Basic, Advanced, and Syntax. Statistical Innovations Inc., Belmont Massachusetts.