Joint modelling of choice and rating data: theory and examples

Edward JD Webb* Stephane Hess†

July 7, 2021

Abstract

In many cases, ordinal data, for example rating objects on a scale from 1 to 5, is observed only for those objects that have been chosen from a set of discrete alternatives, with no ratings for unchosen objects. An example is customer ratings of goods sold by online retailers. The joint modelling of choice and rating is made difficult by the missing ratings for unchosen alternatives. A method of jointly modelling choice and rating data termed a choice-ordered logit (COL) model is presented. Two types of COL model are defined: two-step, which places a positive probability on the chosen alternative not having the highest rating, and one-step, where the highest rated alternative is always chosen. Three case studies exemplifying the use of COL models are given. One uses simulated data and two use data from discrete choice experiments. It is shown that COL models can produce robust estimates. Two-step models provided a better fit than one-step, and most participants seemed to use two-step decision-making. However, a sizeable minority used one-step decision-making in one case study. It is argued that COL models have benefits over standard approaches, in particular adding information on strength-of-preference to discrete choices.

Keywords: discrete choice; ratings; ordered logit; stated preference; joint modelling

*Academic Unit of Health Economics, Leeds Institute of Health Sciences, University of Leeds
†Choice Modelling Centre, University of Leeds

1 Introduction

There are many sources of data on individuals rating objects on an ordinal scale. Often though, rating occurs only after individuals have made a discrete choice between several alternatives. This means that ratings of unchosen alternatives are not observed. An example of this arises in the case of retail websites such as Amazon, where ratings of between one and five stars are only observed for the goods consumers chose to purchase. Other non-retail websites allow ratings of goods purchased elsewhere, such as IMDB for films or Yelp for restaurants and other services. Here, again, cinema goers must first choose what film they want to see, and diners must choose which restaurant to eat at, and their ratings for non-chosen alternatives are not observed. Another example is that of extending stated choice (SC) surveys otherwise known as discrete choice experiments (DCEs), so participants additionally rate their preferred option after each choice.

When decision-makers only rate objects that were previously chosen, the full decision-making process is a joint one. Thus ordered logit (OL) models, which use only rating data, do not fully model the decision-making environment, nor do they make use of the fact that individuals clearly like the objects they are rating enough to select them over others. Likewise, modelling only discrete choices neglects additional information on individuals' preferences provided by the ratings.

This paper introduces a framework for jointly modelling both discrete choices and ratings, which we refer to as choice-ordered logit (COL). It thus captures both components of the decision-making process in a single model, which it is hoped will aid a greater understanding of the whole decision-making environment, both theoretically and empirically. At present, the modelling framework is mostly applicable to stated preference data, although there is a discussion of how it might be extended to reveled preference data in section 4.

This work draws on an extensive literature on modelling both discrete choices and ordered data using both revealed and stated preferences (see Train (2009) and Hess and Daly (2014) for a partial overview). However, we are not aware of any previous studies which combine

choices and ratings in the way proposed here. For example, while some studies have examined how and why individuals leave online ratings (e.g. Moe & Schweidel, 2012) or how consumers react to them (e.g. Sun, 2012), these studies do not explicitly consider the prior discrete choice.

There are several studies in which a DCE is expanded with a Likert scale question, for example Regier, Watson, Burnett, and Ungar (2014) in health, Beck, Fifer, and Rose (2016) in transport, and Mattmann, Logar, and Brouwer (2019) in environmental economics. Yet the Likert scales in these cases measure the uncertainty respondents have over their decisions. Thus they measure something fundamentally different than the strength of preferences for the choice objects, meaning approaches to jointly modelling this data are also conceptually different to the approach detailed here. Here, it is assumed that both discrete choices and ratings are reflective of intrinsic preferences for choice objects.

Gutknecht, Schaarschmidt, Danner, Blome, and Augustin (2018) and Wijnen et al. (2015) conduct DCEs while also measuring how important each attribute was using Likert scales. However, in neither study was the data from both exercises modelled jointly, instead the results of separate models were compared.

Some similarities to the current approach are seen in Rose, Beck, and Hensher (2015), Beck, Rose, and Hensher (2013) and Hensher and Rose (2012). In the survey in those studies, DCE participants were asked whether each item was acceptable or not, which could be thought of as a binary rating which reflects their preferences for choice objects. However, in that case the additional question was asked for each choice object, whereas the focus here is on situations in which only ratings for the chosen object are observed. The additional data was also used to estimate participants' consideration sets, whereas this heuristic is not modelled here.

The remainder of this paper proceeds as follows: Section 2 gives the mathematical framework for the model approach. Section 3 then gives three case studies of applying joint choice-ordered logit models. The first uses simulated data, and the other two use data from

discrete choice experiment (DCE) surveys. Section 4 provides a general discussion of the results from each case study, and section 5 concludes.

2 Theory

Let individual n choose from a set of $\mathcal{J}$ items with $|\mathcal{J}| \geq 2$, then rate the selected item on an integer ordinal scale from 1 to $R \geq 3$, with one being worst and R being best. Assume the utility to n from choosing item $j \in \mathcal{J}$ in task t is $U^C_{jnt} = V_{jnt} + \varepsilon_{jnt}$ where ε_{jnt} follows an extreme value distribution across observations and alternatives. Let $V_{jnt} = \boldsymbol{\beta}_n \boldsymbol{x}_{jnt}$ where $\boldsymbol{x}_{jnt}$ is a vector describing the features of option j as faced by individual n in task t (which may include an alternative specific constant), and $\boldsymbol{\beta}_n$ is a vector describing n's preferences/sensitivities for those features. Assume the utility function when rating is $U^R_{jnt} = \zeta V_{jnt} + \eta_{jnt}$ where η_{jnt} follows an extreme value distribution and ζ is an (optional) scale parameter.

With a type I extreme value distribution for the error terms ε_{jnt}, we obtain a multinomial logit (MNL) model, such that the probability of individual n choosing option i in task t is given by:

$$P^C_{nt}\left(i \mid \boldsymbol{\beta}_n, S_{nt}\right) = \frac{e^{V_{int}}}{\sum_{j \in \mathcal{S}} e^{V_{jnt}}}, \tag{1}$$

where S_{nt} is n's choice set in task t and $\mathcal{S}$ is the universal choice set, i.e. $\mathcal{S} = \langle 1, \ldots, J \rangle$.

The probability of item i being rated r is given by:

$$\begin{aligned} P^R_{int}\left(r \mid \boldsymbol{\beta}_n, \zeta\right) &= P\left(\zeta V_{int} + \epsilon_{int} < \tau_r\right) - P\left(\zeta V_{int} + \epsilon_{int} < \tau_{r-1}\right) \\ &= \frac{1}{1 + e^{\zeta V_{int} - \tau_r}} - \frac{1}{1 + e^{\zeta V_{int} - \tau_{r-1}}} \end{aligned} \tag{2}$$

where $\tau_0, \tau_1, \ldots, \tau_R$ are a series of thresholds defining the utility thresholds at which a given rating is observed, with $\tau_{r-1} < \tau_r$, $\tau_0 = -\infty$ and $\tau_R = \infty$. This probability is of the ordered logit (OL) type.

COL models may follow a one or two-step decision process. We will now look at these two possibilities in turn, where we define $choice_{nt}$ to be the alternative chosen by n in task t, and let $rate_{i_{nt}}$ be the rating given to that alternative.

With a two-step process, choices and ratings are separate, so that there is a possibility that individuals choose an option that does not have the highest rating. The joint probability of choosing item i_{nt}, i.e. $choice_{nt} = i_{nt}$ and rating it r, i.e. $rate_{i_{nt}} = r_{i_{nt}}$ is then:

$$P_{nt}\left(i_{nt}\,\&\,r_{i_{nt}} \mid \boldsymbol{\beta}_n, \zeta\right) = P_{nt}^{C}\left(i_{nt} \mid \boldsymbol{\beta}_n, S_{nt}\right) \cdot P_{int}^{R}\left(r_{i_{nt}} \mid \boldsymbol{\beta}_n, \zeta\right). \tag{3}$$

With a one-step decision process, individuals always choose the item with (weakly) the highest rating. We have previously introduced $\mathcal{S}$ as the universal choice set. Let us now define $\mathcal{S}_{u,nt}$ as the set of unchosen alternatives for individual n in task t, i.e. $\mathcal{S}_{u,nt} = \mathcal{S} \setminus i_{nt}$, and $U\left(\mathcal{S}_{u,nt}, i_{nt}\right) = \mathcal{S}$.

Let $\mathcal{S}_{u,nt}^{\ell m}$ denote the ℓ^{th} (unique) subset of $\mathcal{S}_{u,nt}$ containing exactly m elements, with $m = 0, 1, \ldots, J-1$ and with $\ell = 1, 2, \ldots$ up to ${}_{J-1}C_m$ which is the number of ways of selecting m objects from $J-1$. Note that this notation implies that $\mathcal{S}_{u,nt}^{10}$ is an empty subset, and as ${}_{J-1}C_0 = 1 \;\forall\; J$ it is unique. The probability of choosing i_{nt}, i.e. $choice_{nt} = i_{nt}$ and rating it r, i.e. $rate_{i_{nt}} = r_{i_{nt}}$ is then:

$$\begin{aligned} P_{nt}\left(i_{nt}\,\&\,r_{i_{nt}} \mid \boldsymbol{\beta}_n, \zeta\right) = \sum_{m=0}^{J-1} \sum_{\ell=1}^{{}_{J-1}C_m} \Bigg[& \left(\prod_{k \in \mathcal{S}_{u,nt}^{\ell m} \cup \{i_{nt}\}} P_{knt}^{R}\left(r_{i_{nt}} \mid \boldsymbol{\beta}_n, \zeta\right) \right) \\ & \cdot \left(\prod_{k \in \left\{\mathcal{S}_{u,nt}^{\ell m} \cup \{i_{nt}\}\right\}^{c}} \left(\sum_{q=1}^{r_{i_{nt}}-1} P_{knt}^{R}\left(q \mid \boldsymbol{\beta}_n, \zeta\right) \right) \right) \\ & \cdot P_{nt}^{C}\left(i_{nt} \mid \boldsymbol{\beta}_n, \mathcal{S}_{u,nt}^{\ell m} \cup \{i_{nt}\}\right) \Bigg]. \end{aligned} \tag{4}$$

The first bracketed term represents the probability that i_{nt} and a given subset of $m-1$ other items are rated $r_{i_{nt}}$, and the second term represents the probability that all other objects are rated strictly less than $r_{i_{nt}}$. The final term is the probability that i_{nt} is chosen from the m

objects rated exactly $r_{i_{nt}}$. Only those combinations where no alternatives are rated higher than $r_{i_{nt}}$ are included, as, in contrast to a two-step model, the probability of choosing i_{nt} would be zero if any other alternative obtained a higher rating.

Thus, for example, if the individual is choosing between two objects, say 1 and 2, the probability of choosing 1 and rating it as 3 out of 5 is:

$$\begin{aligned} P_{nt}\left(1 \,\&\, 3 \mid \boldsymbol{\beta}_n, \zeta\right) = & P_{1nt}^{R}\left(3 \mid \boldsymbol{\beta}_n, \zeta\right) \\ & \cdot \Bigg(\sum_{q=1}^{2} P_{2nt}^{R}\left(q \mid \boldsymbol{\beta}_n, \zeta\right) \\ & + P_{2nt}^{R}\left(3 \mid \boldsymbol{\beta}_n, \zeta\right) P_{nt}^{C}\left(1 \mid \boldsymbol{\beta}_n, \mathcal{S}_{12}\right) \Bigg). \end{aligned} \tag{5}$$

The first line relates to the probability of the chosen alternative (i.e. option 1) being rated 3. In the second line, we cover all cases where the other alternative is rated strictly less than 3, meaning that the probability of choosing alternative 1 is equal to 1 as its rating is strictly greater than that for any others. The third line, covers the case where both alternatives are rated as 3, meaning that the discrete choice over two alternatives comes into play, with $\mathcal{S}_{12} = \{1, 2\}$. Any cases where alternative 2 is rated more than 3 do not contribute to the probability of the observed outcome, given that in a one-step model the unchosen alternative could never be rated higher than the chosen alternative.

With three alternatives, and again assuming that alternative 1 is chosen and given a

rating of 3, we simply have more combinations of cases, namely:

$$
\begin{aligned}
P_{nt}\left(1 \,\&\, 3 \mid \boldsymbol{\beta}_n, \zeta\right) = & P_{1nt}^{R}\left(3 \mid \boldsymbol{\beta}_n, \zeta\right) \\
& \cdot \Bigg[\left(\sum_{q=1}^{2} P_{2nt}^{R}\left(q \mid \boldsymbol{\beta}_n, \zeta\right) \right) \cdot \left(\sum_{q=1}^{2} P_{3nt}^{R}\left(q \mid \boldsymbol{\beta}_n, \zeta\right) \right) \\
& + P_{2nt}^{R}\left(3 \mid \boldsymbol{\beta}_n, \zeta\right) \cdot \left(\sum_{q=1}^{2} P_{3nt}^{R}\left(q \mid \boldsymbol{\beta}_n, \zeta\right) \right) \cdot P_{nt}^{C}\left(1 \mid \boldsymbol{\beta}_n, \mathcal{S}_{12}\right) \\
& + \left(\sum_{q=1}^{2} P_{2nt}^{R}\left(q \mid \boldsymbol{\beta}_n, \zeta\right) \right) \cdot P_{3nt}^{R}\left(3 \mid \boldsymbol{\beta}_n, \zeta\right) \cdot P_{nt}^{C}\left(1 \mid \boldsymbol{\beta}_n, \mathcal{S}_{13}\right) \\
& + P_{2nt}^{R}\left(3 \mid \boldsymbol{\beta}_n, \zeta\right) \cdot P_{3nt}^{R}\left(3 \mid \boldsymbol{\beta}_n, \zeta\right) \cdot P_{nt}^{C}\left(1 \mid \boldsymbol{\beta}_n, \mathcal{S}_{123}\right) \Bigg].
\end{aligned}
$$

Here the top line again represents the probability that alternative 1 is rated 3. The next line represents the probability that alternative 1 "wins" outright in the ratings, i.e. alternatives 2 and 3 are rated strictly less than 3. The third line represents the probability of a two-way "tie", i.e. alternative 2 is also rated 3, and alternative 1 is chosen from the two equal highest ranked alternatives. Likewise, the fourth line represents the probability of a two-way tie between alternatives 1 and 3. The final line gives the probability of all three alternatives being ranked 3, and alternative 1 being chosen from among the three.

3 Case studies

3.1 Case study 1

3.1.1 Data

The aim of this case study was to see whether COL models were capable of recovering parameters when the underlying data generating process (DGP) was known. It uses synthetic data which simulates choice situations in which individuals are presented with a binary choice between items characterised by two attributes, then rating their selected item on a scale from

1 to 5, with 1 being worst and 5 being best. The choice situations that simulated participants responded to were created using NGene[1] to generate a Bayesian D-efficient DCE design with five blocks of 10 questions each. Dominated choices were ruled out and the priors for β_1 and β_2 were set to 10^{-4} and -10^{-4}, respectively.

Individual n's utility for choosing item i in task t was defined as

$$U_{int}^{C} = \alpha_i + \beta_1 x_{int,1} + \beta_2 x_{int,2} + \varepsilon_{int} \tag{6}$$

where $x_{int,k} \in \{1, 2, 3, 4\}$ gives the level of attribute $k \in \{1, 2\}$ for object i, β_k represents n's choice preferences for attribute k and α_i is an alternative specific constant. For all individuals β_1 was set to 0.9 and β_2 was set to -1.1. Thus attribute 1 represents a desirable feature (e.g. quality) and attribute 2 represents an undesirable feature (e.g. price). The ASCs were set to $\alpha_1 = 0.4$, $\alpha_2 = 0$. The error term ε_{int} followed a type I extreme value distribution with a variance of $\pi^2/6$, implying that we obtain an MNL model.

Individual n's preferences for rating attribute i in task t are similarly given by

$$U_{int}^{R} = \zeta \left(\alpha_i + \beta_1 x_{int,1} + \beta_2 x_{int,2}\right) + \eta_{int} \tag{7}$$

with i being rated r if $\tau_{r-1} < U_{int}^{R} < \tau_r$. For all individuals, the thresholds were set to $\tau_1 = -3$, $\tau_2 = -1$, $\tau_3 = 1$, $\tau_4 = 3$. In a given simulation, the same value of ζ was chosen for each person, but different values were used in different simulations. Six datasets were generated, each with 500 simulated responses. Three datasets had a two-step DGP with values of 0.5, 1 and 2 chosen for ζ. The other three datasets had a one-step DGP, with again values of 0.5, 1 and 2 chosen for ζ.

In addition, a series of further datasets were derived by mixing between datasets with the same scale parameter but different decision steps to mimic the case where there is heterogeneity in individuals' decision processes. The mixed datasets also had had 500 synthetic

[1] www.choice-metrics.com

respondents with a fraction drawn from a two-step and the remainder drawn from the equivalent one-step database. For each pair of two-step and one-step datasets, between 1% and 99% of respondents were randomly drawn from the two-step datasets, in steps of one percentage point.

3.1.2 Analysis

For each dataset, the following models were estimated:

(i) MNL;

(ii) OL;

(iii) Two-step COL without scale parameter;

(iv) One-step COL without scale parameter;

(v) Two-step COL with scale parameter;

(vi) One-step COL with scale parameter.

For the mixed datasets, two-step and one-step models were estimated with and without a scale parameter. Whether estimated parameters were significantly different from the true underlying values at the 5% level was assessed using t-tests, with adjustment for multiple testing using Holm's sequential Bonferonni correction (Holm, 1979). The results of two- and one-step estimation were then used in a model averaging process (Hancock & Hess, 2020) to estimate the fraction of individuals using a given decision process. (This can be thought of as estimating a finite-mixture model with heterogeneity in the decision-making process, with the parameters within a given decision-making class fixed at the values from models estimated without mixing.) All estimation was performed using the Apollo choice modelling package for R (Hess & Palma, 2019).

3.1.3 Results

Table 1 gives the results of model estimation. It can be seen that in most cases COL models return the true taste parameters of the synthetic respondents with a reasonable degree of accuracy. Even where significant differences were observed between estimated and true taste parameters, the estimated and true marginal rates of substitution are often similar. With a two-step DGP, MNL and OL performed well, however with a one-step DGP, MNL and OL often performed worse than a one-step COL model, especially when the true underlying scale parameter differed from 1. There was variation in how often the true rating utility thresholds were returned. Estimated parameters were significantly different when COL models were estimated with a different number of decision steps compared to the true DGP, as well as when the COL model and DGP were both two-step, but the COL model had no scale parameter and the DGP had a true scale parameter of 2.

Adding a scale parameter to COL models often improved accuracy. Notable exceptions are one-step COL models with a scale parameter when the DGP was two-step with a non-unity scale parameter. When the true scale parameter was 2, the one-step COL model estimated very small taste parameters ($\hat{\alpha}_1 = 0.02$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = -0.06$) and a large scale parameter ($\hat{\zeta} = 33.6$). This may be interpreted as the best fit resulting from the model only explaining the rating data, while the choice data is modelled as being largely random. In general when the decision process in the model and DGP were mismatched, two-step COL models performed better with a one-step DGP than one-step models with a two-step DGP.

Figure 1 illustrates the BIC of the estimated models, including for MNL and OL models combined. Although differences may be small, a correctly specified COL model provided a better fit than combined MNL and OL in every case. Sometimes even misspecified COL models performed better, for example with a one-step DGP with scale parameter 1, combined MNL and OL provided the worst fit of all models, although again differences were sometimes small.

Figure 2 summarises the results of model averaging. In general it was possible to recover

Table 1: Case study 1 model results

True DGP	Estimated model	α_1		β_1		β_2		ζ		τ_1		τ_2		τ_3		τ_4	
2-step	MNL	0.379	(0.037)	0.959	(0.031)	-1.16	(0.035)										
$\zeta = 0.5$	OL	0.220*	(0.053)	0.465*	(0.03)	-0.571*	(0.033)			-3.02	(0.094)	-1	(0.072)	1.06	(0.076)	3.12	(0.093)
	COL 2-step no scale	0.304*	(0.028)	0.696*	(0.021)	-0.856*	(0.022)			-3.11	(0.078)	-1.05	(0.049)	1.11	(0.051)	3.22	(0.077)
	COL 1-step no scale	0.359	(0.034)	0.687*	(0.023)	-0.865*	(0.024)			-1.64*	(0.066)	-0.154*	(0.052)	1.70*	(0.055)	3.72*	(0.08)
	COL 2-step scale	0.385	(0.035)	0.957	(0.031)	-1.16	(0.035)	0.491	(0.029)	-3.02	(0.072)	-1.01	(0.04)	1.05	(0.04)	3.11	(0.072)
	COL 1-step scale	0.557	(0.065)	1.14	(0.099)	-1.42	(0.119)	0.564	(0.054)	-1.61*	(0.062)	-0.143*	(0.048)	1.69*	(0.051)	3.70*	(0.077)
1-step	MNL	0.219*	(0.032)	0.548*	(0.027)	-0.664*	(0.029)										
$\zeta = 0.5$	OL	-0.0223*	(0.053)	0.387*	(0.029)	-0.391*	(0.03)			-6.01*	(0.298)	-2.21*	(0.076)	0.309*	(0.069)	2.40*	(0.081)
	COL 2-step no scale	0.155*	(0.027)	0.463*	(0.021)	-0.546*	(0.021)			-6.07*	(0.279)	-2.31*	(0.06)	0.247*	(0.045)	2.36*	(0.063)
	COL 1-step no scale	0.224*	(0.033)	0.535*	(0.022)	-0.651*	(0.022)			-3.37	(0.167)	-1.05	(0.057)	0.993	(0.05)	2.98	(0.067)
	COL 2-step scale	0.192*	(0.031)	0.563*	(0.026)	-0.666*	(0.028)	0.601	(0.047)	-6.03*	(0.277)	-2.27*	(0.054)	0.240*	(0.037)	2.32*	(0.057)
	COL 1-step scale	0.415	(0.074)	1.03	(0.116)	-1.25	(0.141)	0.455	(0.062)	-3.22	(0.145)	-1.01	(0.052)	0.987	(0.044)	2.96	(0.062)
2-step	MNL	0.331	(0.034)	0.921	(0.03)	-1.14	(0.033)										
$\zeta = 1$	OL	0.405	(0.05)	0.876	(0.033)	-1.1	(0.033)			-3.08	(0.094)	-1	(0.073)	1.02	(0.077)	3.05	(0.091)
	COL 2-step no scale	0.351	(0.027)	0.899	(0.022)	-1.12	(0.023)			-3.11	(0.072)	-1.02	(0.048)	1.01	(0.05)	3.05	(0.069)
	COL 1-step no scale	0.385	(0.034)	0.861	(0.024)	-1.1	(0.025)			-1.71*	(0.062)	-0.195*	(0.052)	1.53*	(0.053)	3.44*	(0.071)
	COL 2-step scale	0.356	(0.028)	0.916	(0.029)	-1.14	(0.033)	0.964	(0.041)	-3.1	(0.072)	-1.02	(0.048)	1	(0.049)	3.04	(0.07)
	COL 1-step scale	0.172*	(0.03)	0.354*	(0.06)	-0.456*	(0.076)	2.63*	(0.458)	-1.78*	(0.067)	-0.206*	(0.056)	1.55*	(0.058)	3.48*	(0.075)
1-step	MNL	0.306	(0.033)	0.842	(0.028)	-1.03	(0.031)										
$\zeta = 1$	OL	0.0683*	(0.054)	0.823	(0.032)	-0.930*	(0.033)			-4.62*	(0.14)	-1.94*	(0.08)	0.347*	(0.077)	2.43*	(0.086)
	COL 2-step no scale	0.238*	(0.029)	0.823*	(0.022)	-0.987*	(0.024)			-4.67*	(0.127)	-1.99*	(0.054)	0.316*	(0.05)	2.42*	(0.063)
	COL 1-step no scale	0.346	(0.036)	0.941	(0.024)	-1.16	(0.025)			-2.92	(0.09)	-1.02	(0.054)	0.961	(0.053)	2.96	(0.065)
	COL 2-step scale	0.250*	(0.031)	0.861	(0.027)	-1.03	(0.03)	0.91	(0.036)	-4.64*	(0.127)	-1.97*	(0.053)	0.313*	(0.048)	2.39*	(0.062)
	COL 1-step scale	0.383	(0.053)	1.04	(0.095)	-1.28	(0.117)	0.898	(0.092)	-2.91	(0.09)	-1.02	(0.054)	0.959	(0.052)	2.96	(0.065)
2-step	MNL	0.387	(0.035)	0.9	(0.028)	-1.1	(0.031)										
$\zeta = 2$	OL	0.706*	(0.056)	1.82*	(0.038)	-2.18*	(0.044)			-2.95	(0.091)	-0.957	(0.075)	0.986	(0.072)	3.01	(0.087)
	COL 2-step no scale	0.540*	(0.034)	1.34*	(0.025)	-1.62*	(0.03)			-2.61*	(0.066)	-0.855	(0.053)	0.827*	(0.049)	2.59*	(0.058)
	COL 1-step no scale	0.582*	(0.044)	1.28*	(0.028)	-1.58*	(0.033)			-1.44*	(0.07)	-0.126*	(0.061)	1.30*	(0.058)	2.89	(0.064)
	COL 2-step scale	0.367	(0.024)	0.909	(0.026)	-1.1	(0.031)	1.99	(0.065)	-2.96	(0.082)	-0.967	(0.065)	0.977	(0.061)	3	(0.075)
	COL 1-step scale	0.0223*	(0.003)	0.0471*	(0.004)	-0.0580*	(0.005)	33.6*	(2.913)	-1.65*	(0.076)	-0.129*	(0.065)	1.45*	(0.062)	3.18	(0.07)
1-step	MNL	0.661*	(0.042)	1.41*	(0.038)	-1.74*	(0.045)										
$\zeta = 2$	OL	0.591*	(0.056)	1.69*	(0.038)	-2.02*	(0.044)			-3.98*	(0.109)	-1.69*	(0.088)	0.438*	(0.08)	2.47*	(0.085)
	COL 2-step no scale	0.663*	(0.038)	1.56*	(0.031)	-1.90*	(0.036)			-3.92*	(0.089)	-1.68*	(0.066)	0.389*	(0.057)	2.37*	(0.065)
	COL 1-step no scale	0.813*	(0.043)	1.71*	(0.031)	-2.12*	(0.035)			-2.95	(0.081)	-1.04	(0.067)	0.898	(0.06)	2.83	(0.068)
	COL 2-step scale	0.597*	(0.037)	1.43*	(0.037)	-1.73*	(0.044)	1.16*	(0.033)	-4.01*	(0.094)	-1.72*	(0.07)	0.413*	(0.06)	2.44*	(0.069)
	COL 1-step scale	0.473	(0.052)	1.03	(0.099)	-1.26	(0.121)	1.74	(0.179)	-3.03	(0.086)	-1.06	(0.069)	0.925	(0.062)	2.89	(0.07)

Note. Standard errors in parentheses; * = significantly different from true underlying parameter at 5% level after adjustment using Holm's sequential Bonferonni correction

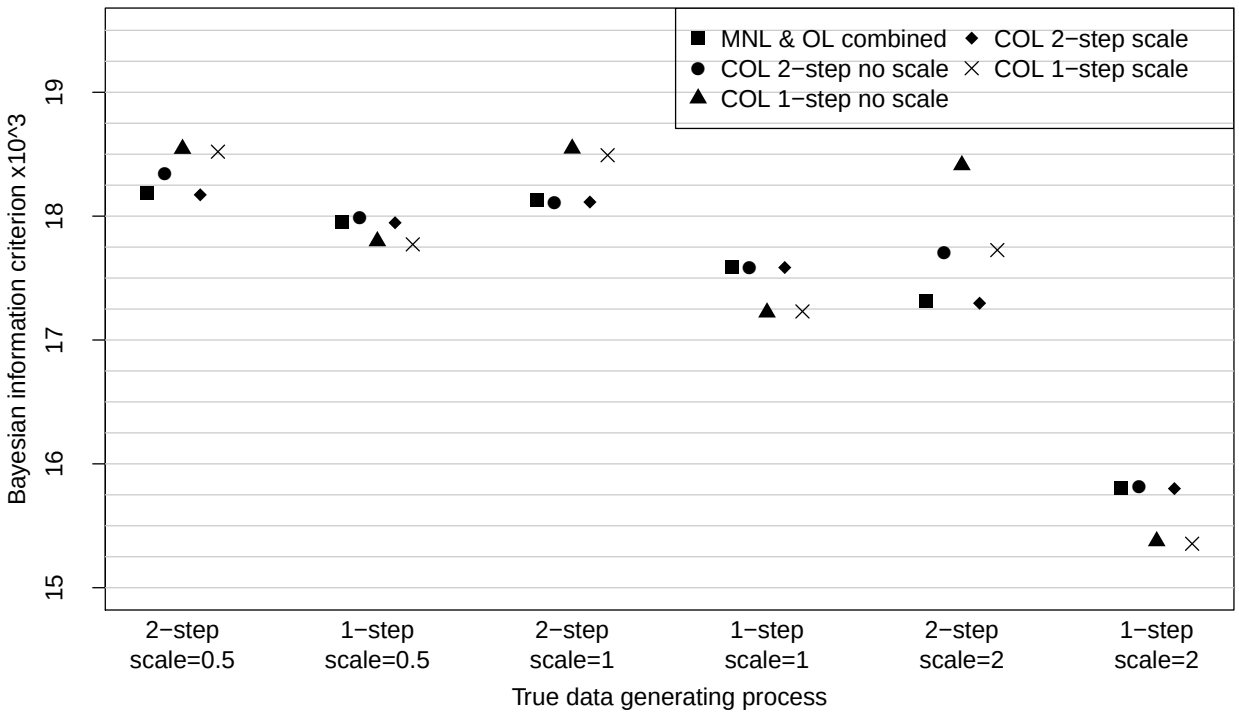


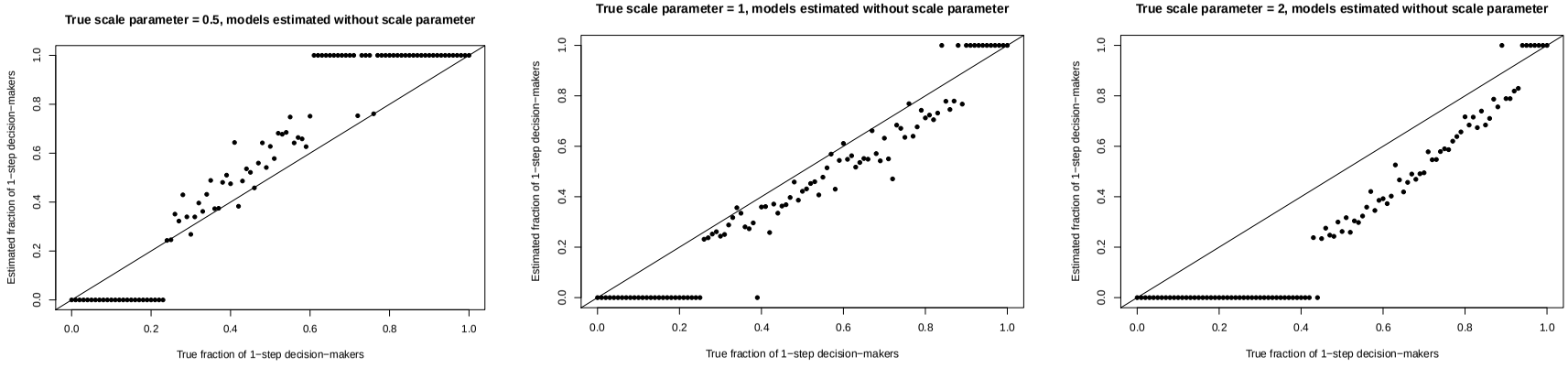
Figure 1: Bayesian information criteria for case study 1 models

the fractions of one-step and two-step decision-makers in the population, as long as the split was not too one-sided. Having a scale parameter in the estimated models generally improved accuracy in the cases where the true underlying parameter was 0.5 or 2.

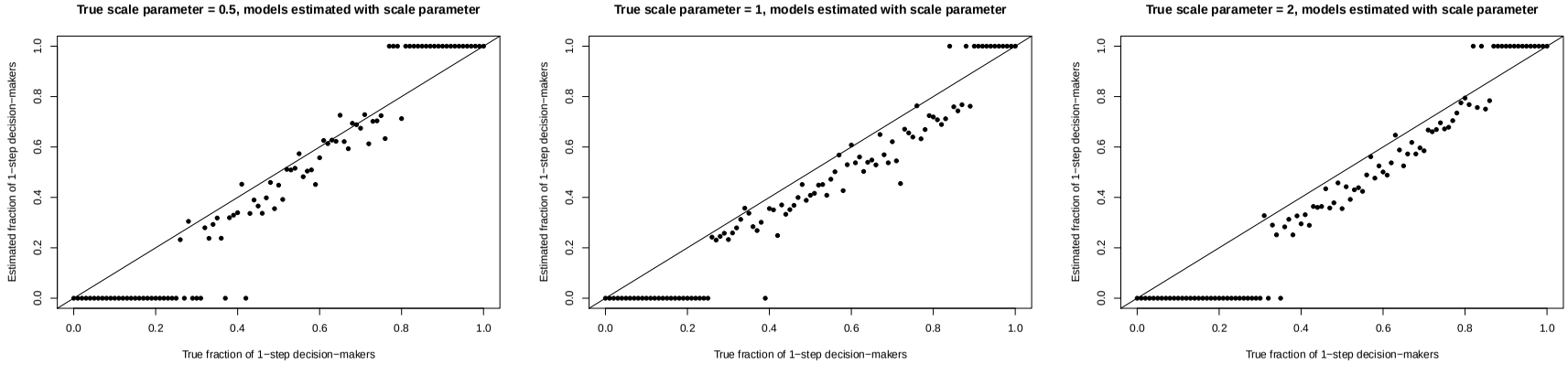
3.1.4 Discussion

This simulation exercise has acted as a proof-of-concept for COL models. It has demonstrated that COL models are capable of accurately estimating the parameters of a known data-generating process, with discrepancies between estimated and true underlying values most often seen when the model and DGP were mis-matched. This means that they can be applied in situations with an unknown data-generating process with some confidence that the resulting estimated model parameters give insight into it.

The results of comparing combined MNL and OL BIC to COL models reveals that in every case there was a COL model that provided a more parsimonious fit. In some cases, it may be that a researcher is only interested in modelling either choices or ratings. However,



(a) (b) (c)



(d) (e) (f)

Figure 2: Results of model averaging for datasets with heterogeneous decision-making

the simulation exercise has shown that there may be an advantage to using COL models even in these cases. In particular when the true DGP was one-step, one-step COL models usually provided more accurate parameter estimates than either MNL or OL models. Thus if the researcher believes individuals are using a one-step decision process, a COL model could be a better choice than modelling only choices or only ratings. This is more likely to be the case with a one-step decision-making process as with it, MNL and OL are misspecified in a way that they are not with a two-step decision-making process.

The ASCs were included in the utility functions for both choices and ratings. This is appropriate when there may be some intrinsic quality of the alternatives, aside from those described by other attributes, which people may have preferences for. For example, individuals may have an intrinsic preference for travelling by train rather than bus, aside from their preferences for time, cost, etc. They would then be more likely to choose to travel by train, and to rate train journeys higher than bus journeys, making a train-specific constant appropriate to include in both choice and rating utility functions. However, in some contexts it could be appropriate only to include the ASCs in the choice utility function. An example is DCEs, where in some cases participants may have a tendency to pick the option presented on the left-hand side. This could be modelled by including an ASC for the left-hand option in the choice utility function. However, it is less likely that the layout would influence the rating of the chosen alternative, so an ASC would be less appropriate to include in the rating utility function.

It is not certain with real choices and rating how individuals will act, and in particular whether their decision-making will be one-step or two-step. The decision-making process may also vary from situation to situation. Hence it is useful to use simulation to compare cases when it is known whether the model is correctly specified or not. In many cases a misspecified COL model still produced reasonably accurate model estimates, particularly a two-step model with a one-step DGP with a scale parameter of 1. However, one-step models could produce imprecise parameter estimates if misspecified, including a case where very

little variation in choices was explained by the model.

There may be heterogeneity amongst people in whether a two-step or a one-step decision-making process is used. Hence it is encouraging that it was possible when the fraction of one type of decision-makers not too large to accurately recover the fraction of individuals using a given process using model averaging. The next two case studies use real data, and in each we investigate what fraction of individuals use two/one-step decision-making. Demonstrating that model averaging can produce accurate estimates in synthesised data where the true underlying fractions are known is important to give confidence in the results with real data.

3.2 Case study 2 - discrete choice experiment on decision-making of alternative and augmentative communication professionals

3.2.1 Data

Alternative and augmentative communication (AAC) describes a large number of techniques which allow people with no natural speech to communicate. AAC systems vary a lot. They can be low-tech, for example boards with letters people can point to, or high tech, for example the speech-synthesising system used by the late Stephen Hawking. People who use AAC have a wide variety of conditions, including cerebral palsy, intellectual/developmental delay and autism spectrum conditions, and even within diagnoses, people's needs and abilities vary significantly (Murray & Goldbart, 2009). This means the features of an AAC system must be carefully matched to the individual needs of a child, otherwise the child may abandon using it (Moorcroft, Scarinci, & Meyer, 2019).

The data for this case study is taken from a DCE examining UK-based AAC professionals' decision-making when choosing AAC systems for children (Webb, Lynch, et al., 2019). In the survey, participants were shown a short vignette describing a hypothetical child described by four attributes: their language skills, their determination, their communication ability with AAC, and their expected future trajectory, with two, three, three and three levels respectively.

They then had to choose between three hypothetical AAC systems for the child. The AAC systems were characterised by five attributes: type and size of pre-provided vocabulary (three levels each), type of vocabulary organisation (four levels), type of graphic symbols used (four levels) and how consistent the layout of the interface is (three levels). After making their choice, they were asked to rate how good a fit their preferred option was for the child, i.e. how well the features of the system suited the child's needs and circumstances, on a scale from 1 (poor fit) to 7 (good fit). An example choice task is shown in Figure 3. Note that the attribute names used above are somewhat simplified for a non-AAC audience; the full list of attributes and levels shown to participants is provided in the appendix for the interested reader.

Experts on AAC removed 18 out of a possible 54 child vignettes and 158 out of a possible 432 AAC systems as being unrealistic. A D-efficient design was generated for AAC systems using NGene with five blocks of 12 choice tasks each. Respondents were randomly allocated to a block as well as being independently randomly allocated three child vignettes, answering four choice tasks for each vignette. Responses were collected online between 20 October 2017 and 4 March 2018, with 155 participants completing the DCE.

3.2.2 Analysis

In Webb, Lynch, et al. (2019) a model was estimated with interactions between AAC systems and child attributes, as the study sought to investigate how participants' preferences changed when choosing for children with different characteristics. However, in the current study for simplicity no interactions were included in models, and coefficients were only included for AAC systems. The same six models as in case study 1 were estimated. For each model, t-tests were used to judge if coefficients were different from 0, with significance judged at the 5% level after adjustment for multiple testing using Holm's sequential Bonferroni correction. Model averaging was performed with pairs of two- and one-step models with and without a scale parameter. From the results, posterior probabilities of participants belonging to the

Child B has delayed expressive and receptive language and able to use aided AAC for a few communicative functions. Child B is only motivated to communicate through methods other than symbol communication systems. Child B is predicted to maintain current skills and abilities (plateau).

	System 1	System 2	System 3
Graphic Representation Primary type of graphic symbol used	Ideographic symbols	Photo symbols	Pictographic symbols
Consistency of layout Consistency of layout of symbols on pages, including when navigating through pages to select desired output.	Highly consistent layout	Somewhat consistent layout	Inconsistent layout
Vocabulary sets Pre-determined vocabulary or language package provided	Commercially provided sets without language progression	Commercially provided sets with language progression	Commercially provided sets without language progression
Size of vocabulary The size of the output vocabulary available within the aided AAC system.	More than 1000 vocabulary items	50-1000 vocabulary items	Up to 50 vocabulary items
Type of vocabulary organisation Primary format used to organise the vocabulary within the aided AAC system	Pragmatic organisation	Semantic syntactic organisation	Visual scene display
For this child I would choose:	○	○	○

On a scale from 1 to 7, how good a match is your chosen device for this child? (1=very unsuitable, 7=very suitable)

1 = very unsuitable ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 = very suitable ○

35%

Figure 3: Example choice task from case study 2 discrete choice task

two-step and one-step decision-making classes were estimated (Hess, 2014). All estimation was performed using the Apollo choice modelling package for R (Hess & Palma, 2019).

3.2.3 Results

The results of model estimation are given in Table 2. With the MNL model, 10 out of 12 taste coefficients were significant, more than in the OL model, which had only four. This could suggest that with information on the characteristics of the chosen option only, it is difficult to explain the rating of that option on the basis of its attributes. This is a key motivation for the use of the COL models. With COL models, a similar number of coefficients (8-9) were significant as with MNL for both two-step models and the one-step model with no scale parameter. Only six parameters were significant in the one-step COL with a scale parameter.

Some parameters had the opposite sign in the MNL and OL models, but in every case the coefficients were insignificant in the OL model, and in two cases in the MNL model as well. Otherwise, coefficients were of similar magnitudes. Coefficients in the two-step COL models and the one-step model with no scale parameter were similar in magnitude to the MNL model, apart from two parameters. For those parameters (graphics levels 2 and 4), some sign reversals were seen, but neither parameter was significant in any of the models. The one-step COL model with a scale parameter had low preference coefficients and a large scale parameter (10.9).

Figure 4 compares models' BIC, including MNL and OL combined. All COL models have a better fit than the combined MNL and OL models, with the two-step COL model without a scale parameter giving the best fit.

Model averaging showed that most participants used a two-step decision-making process, with the share if two-step decision-makers being 80.5% and 77.0% respectively for models with and without a scale parameter. Figure 5 gives density plots for the posterior probabilities of decision-making classes. Relatively little heterogeneity is seen, with most of the probability density being clustered around 80% of two-step decision-makers.

Table 2: Case study 2 model results

	MNL	OL	2-step no scale	1-step no scale	2-step scale	1-step scale
Vocab sets L2	0.247*	0.185	0.216*	0.256*	0.219*	0.0284
	(0.079)	(0.124)	(0.073)	(0.079)	(0.074)	(0.011)
Vocab sets L3	0.553*	0.484*	0.524*	0.617*	0.532*	0.0660*
	(0.086)	(0.147)	(0.079)	(0.094)	(0.081)	(0.019)
Vocab size L2	0.467*	0.147	0.355*	0.436*	0.363*	0.0450*
	(0.081)	(0.135)	(0.072)	(0.08)	(0.077)	(0.015)
Vocab size L3	0.519*	0.575*	0.545*	0.629*	0.550*	0.0692*
	(0.092)	(0.158)	(0.084)	(0.103)	(0.085)	(0.02)
Vocab organisation L2	0.296*	-0.207	0.136	0.172	0.144	0.0158
	(0.106)	(0.133)	(0.083)	(0.101)	(0.085)	(0.011)
Vocab organisation L3	0.256*	-0.0116	0.189	0.271*	0.195	0.0265
	(0.103)	(0.142)	(0.086)	(0.106)	(0.087)	(0.013)
Vocab organisation L4	0.404*	-0.0288	0.268*	0.411*	0.277*	0.0411*
	(0.094)	(0.142)	(0.083)	(0.101)	(0.084)	(0.015)
Graphics L2	0.00453	-0.0504	-0.0226	-0.0103	-0.0201	-0.00311
	(0.095)	(0.139)	(0.079)	(0.093)	(0.081)	(0.009)
Graphics L3	-0.304*	-0.269	-0.300*	-0.361*	-0.304*	-0.0381
	(0.103)	(0.152)	(0.091)	(0.107)	(0.091)	(0.015)
Graphics L4	0.0141	-0.175	-0.0568	-0.0417	-0.0532	-0.00717
	(0.092)	(0.137)	(0.077)	(0.094)	(0.08)	(0.01)
Layout consistency L2	0.594*	0.722*	0.628*	0.727*	0.635*	0.0788*
	(0.079)	(0.143)	(0.074)	(0.085)	(0.076)	(0.022)
Layout consistency L3	0.938*	1.01*	0.956*	1.10*	0.968*	0.118*
	(0.086)	(0.136)	(0.073)	(0.091)	(0.077)	(0.032)
Scale parameter					0.95	10.9*
					(0.114)	(2.957)
τ_1		-3.47	-3.15	0.194	-3.19	0.388
		(0.451)	(0.441)	(0.23)	(0.449)	(0.252)
τ_2		-1.76	-1.44	1	-1.48	1.23
		(0.26)	(0.233)	(0.185)	(0.249)	(0.197)
τ_3		-0.664	-0.348	1.63	-0.39	1.87
		(0.222)	(0.178)	(0.171)	(0.197)	(0.186)
τ_4		0.674	0.987	2.55	0.941	2.82
		(0.222)	(0.166)	(0.17)	(0.194)	(0.189)
τ_5		2.23	2.54	3.84	2.49	4.12
		(0.231)	(0.169)	(0.178)	(0.206)	(0.198)
τ_6		4.59	4.9	6.09	4.85	6.37
		(0.306)	(0.249)	(0.25)	(0.288)	(0.267)
BIC	3717.2	5894.9	9540.4	9596.9	9547.7	9577.3

Note. Standard errors in parentheses; * = significant at 5% level after adjustment using Holm's sequential Bonferonni correction

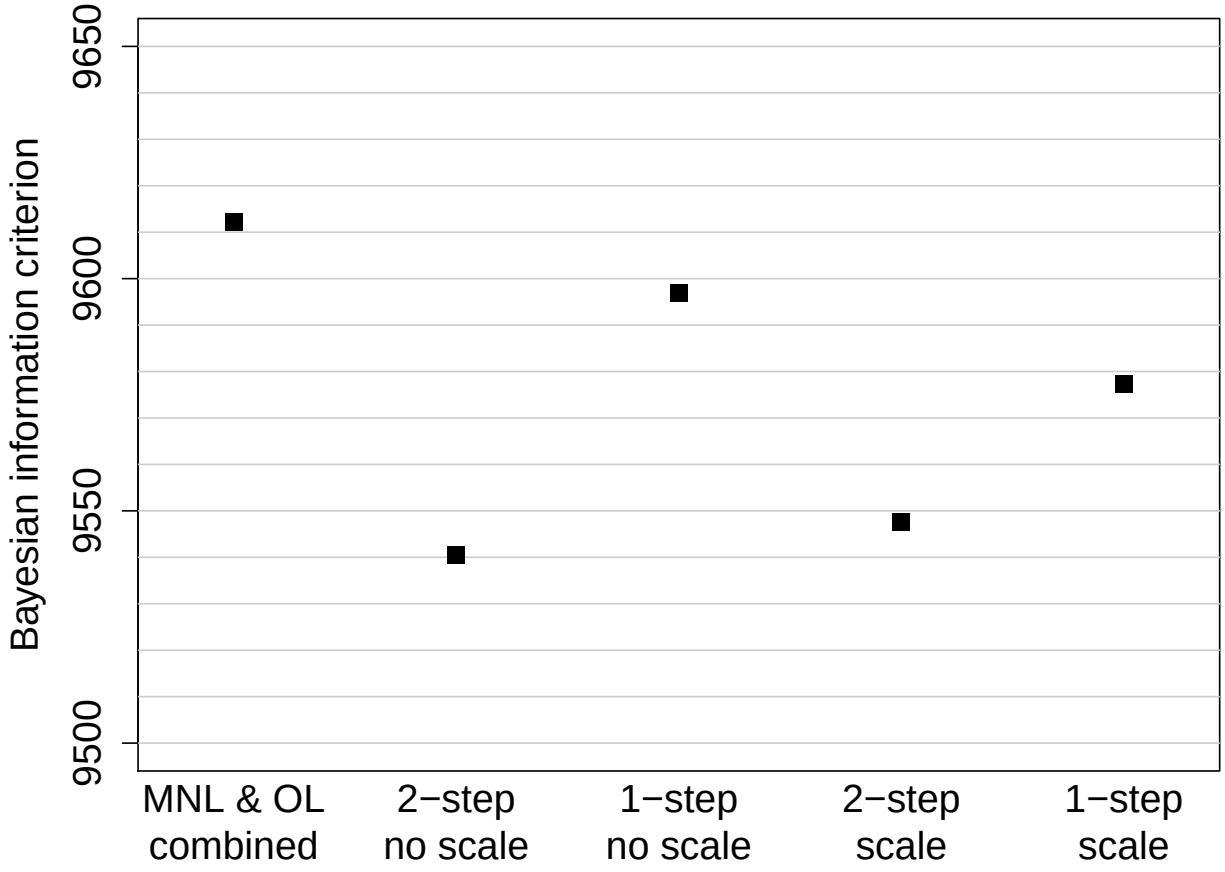
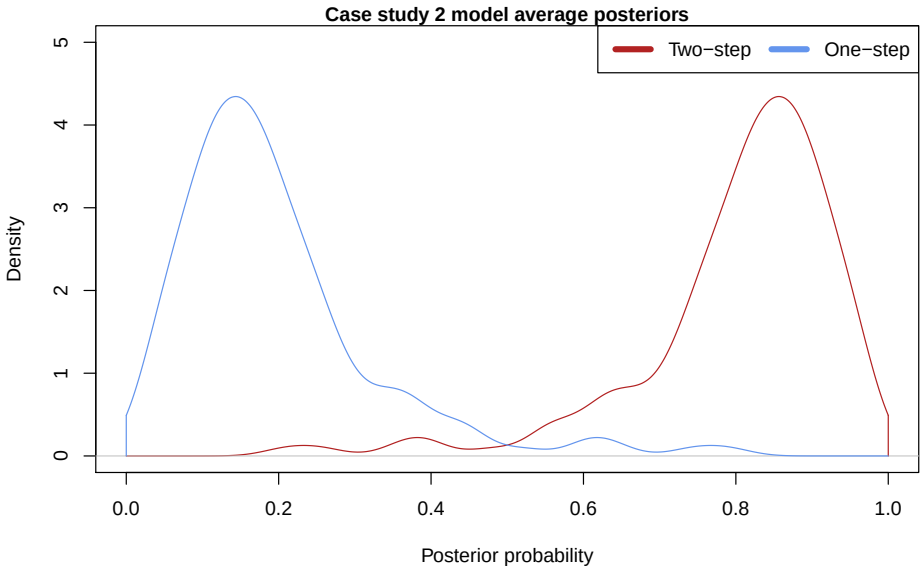
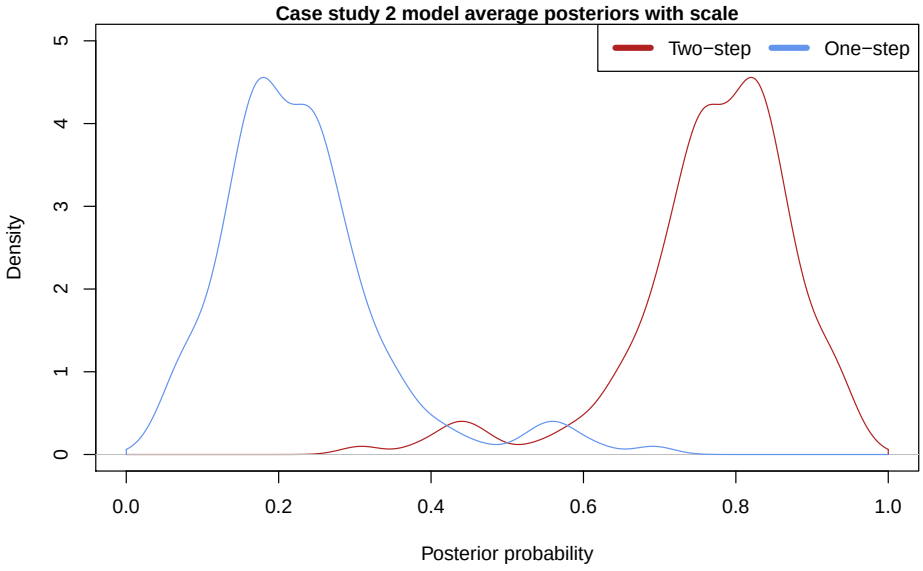


Figure 4: Bayesian information criteria for case study 2 models



(a)



(b)

Figure 5: Model averaging posterior decision-making class allocation for case study 2

3.2.4 Discussion

The dataset for this case study has a relatively low number of respondents, with 155 compared to a median of 401 participants in DCEs in healthcare published from 2013-2017 (Soekhai, de Bekker-Grob, Ellis, & Vass, 2019). However, such a low number was inevitable, as in many other cases, as there are few members of the target population. (It is estimated that there are only around 800 AAC professionals working with children in the UK.[2]) Thus this case study is useful in demonstrating that COL models are feasible when collecting a large amount of data is impractical.

It is encouraging that, apart from the one-step model with a scale parameter, most coefficients are statistically significant, and this also represents a large improvement over the standard OL model. It is also encouraging that, as measured by BIC, COL models provided a better fit than MNL and OL combined.

In this case, it should be expected that participants would use a two-step decision-making process, as there are good reasons to allow a positive probability of participants choosing one AAC system, yet rating it lower than they would have one of the unchosen systems. For example, they may choose a system that is less suitable to child's current needs than alternatives, but which is better for the child to grow into. It could be that respondents were reluctant to choose some AAC systems due to a high inferred cost, which would not impact on how good a fit they considered a system for a child (although in practice previous research has indicated that cost is relatively unimportant in AAC professionals' decision making in the UK (Webb, Meads, et al., 2019)).

The prior expectation that participants would use a two-step decision-making process was borne out by the empirical results, with two-step COL models having lower BIC and the model averaging process estimating around four fifths of participants to be in the two-step decision-making class. A large majority of two-step decision-makers also explains the poor performance of the one-step COL model with a scale parameter. The results were similar

[2]Personal correspondence with Communication Matters, a UK-wide AAC charity.

to the analogous models in case study 1 when estimated on a two-step DGP with a true scale parameter of 2 (see Table 1). The misspecification of the model results in low taste parameters and a high scale parameter, indicating that very little variation in choices is explained.

There is also a conceptual advantage in this case of supplementing discrete choice data with ratings data. Participants made choices between AAC systems for a number of different children. Their discrete choices inform about the relative trade-offs participants made between attributes of AAC systems, and how these relative trade-offs shift in response to changes in child characteristics. However, it is only when this data is combined with the ratings that comparisons on the same scale between different child vignettes can be made. This then allows new insight to be drawn, and new policy relevant research questions to be answered. Examples are whether AAC systems exist for all children that are regarded as a good fit, and whether for some child vignettes there are no highly rated systems.

3.3 Case study 3 - Discrete choice experiment to value EQ-5D-5L

3.3.1 Data

EQ-5D is a generic multi-dimensional measure of health-related quality of life (HrQoL) (Dolan, 1997). It measures HrQoL on five dimensions: mobility, self-care, usual activities, pain or discomfort, and anxiety or depression. There are three-level and five-level versions of the survey. In the five-level version (EQ-5D-5L) respondents indicate for each dimension which one of five statements most closely matches their health that day, with level 1 representing no problems and level 5 representing severe problems.

EQ-5D is often used in economic evaluations of health treatments and services. By assigning a value to each EQ-5D state (a "value set") on a scale with 1 defined as "full health" (level 1 on each dimension) and 0 defined as dead, measure of the HrQoL benefit of a treatment can be constructed. Cost-utility analysis using EQ-5D data as a measure of benefit is the method preferred by the National Institute for Health and Care Excellence

(NICE) in the UK when deciding whether or not to recommend treatments are funded by the NHS.

Recently DCEs have been a popular method to create EQ-5D value sets, e.g. Devlin, Shah, Feng, Mulhern, and van Hout (2018), Mulhern, Bansback, Hole, and Tsuchiya (2017), Ramos-Goñi et al. (2017), Bansback, Brazier, Tsuchiya, and Anis (2012), Stolk, Oppe, Scalone, and Krabbe (2010), and the data from this study comes from a similar exercise to create a value set for the UK. DCE results are on a latent scale, and hence external information is needed to anchor values to the full health=1, dead=0 scale (here a visual analogue scale exercise was used). However, this process of anchoring valuations is not relevant to the current study, thus only relative values are presented.

Survey participants were shown two EQ-5D-5L health states, and asked to choose which they considered to be the best. They then rated their preferred health-state on a scale from 0 (the worst health the respondent can imagine) to 10 (the best health the respondent can imagine). An example choice task is shown in Figure 6.

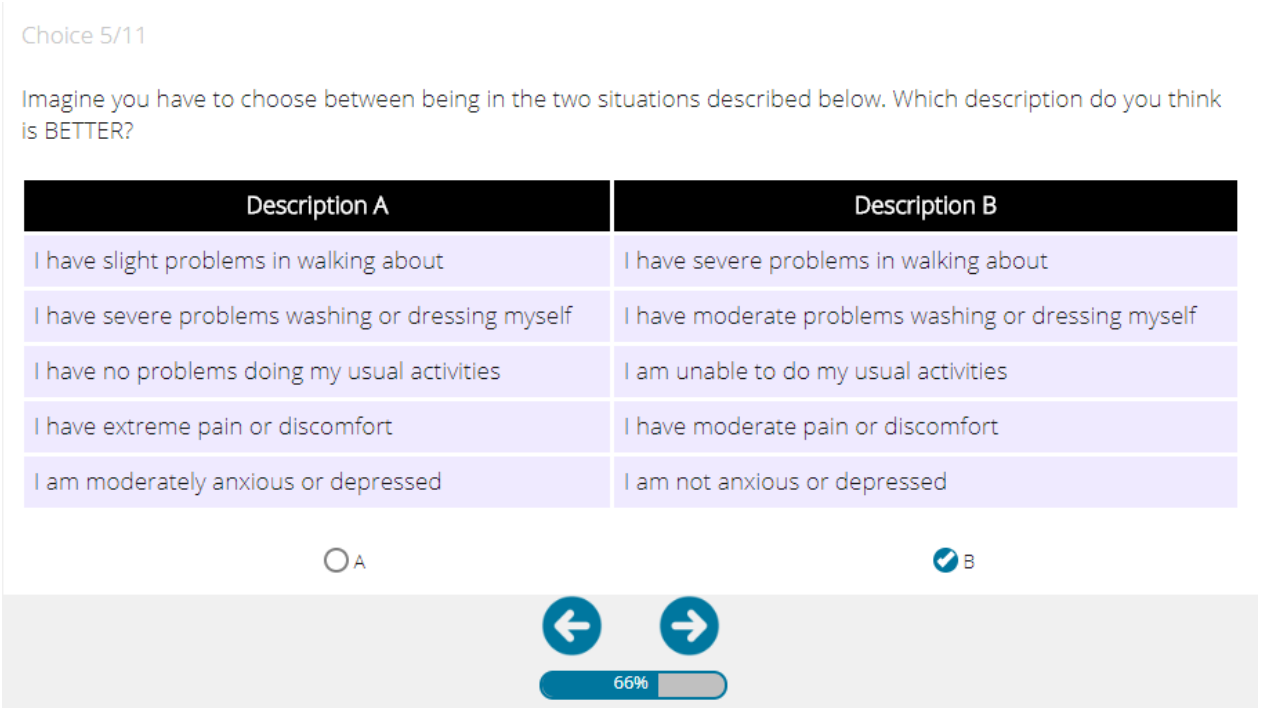
A D-efficient design was created using NGene with 8 blocks with 10 questions each. Two samples were collected, a main sample of 3400 respondents which were representative of the UK general public, and a "boost" sample of 507 UK men aged over 50.[3]

3.3.2 Analysis

Level 1 was chosen as the baseline for each dimension, so that the coefficients on levels 2-5 represent decrements to full health. As participants should value worse health states lower, a logical ordering is imposed that coefficients for higher levels should be more negative than coefficients for lower levels. The same six models as in case studies 1 and 2 were estimated on the main sample. For level 2 coefficients, t-tests were used to assess if they were statistically different from 0. For level 3-5 coefficients, Welch's t-tests CITE were used to assess if they were statistically different from the parameter one level lower (e.g. if a level 3 coefficient was

[3]This demographic was chosen to satisfy the aims of a separate research project

(a) Discrete choice



(b) Rating

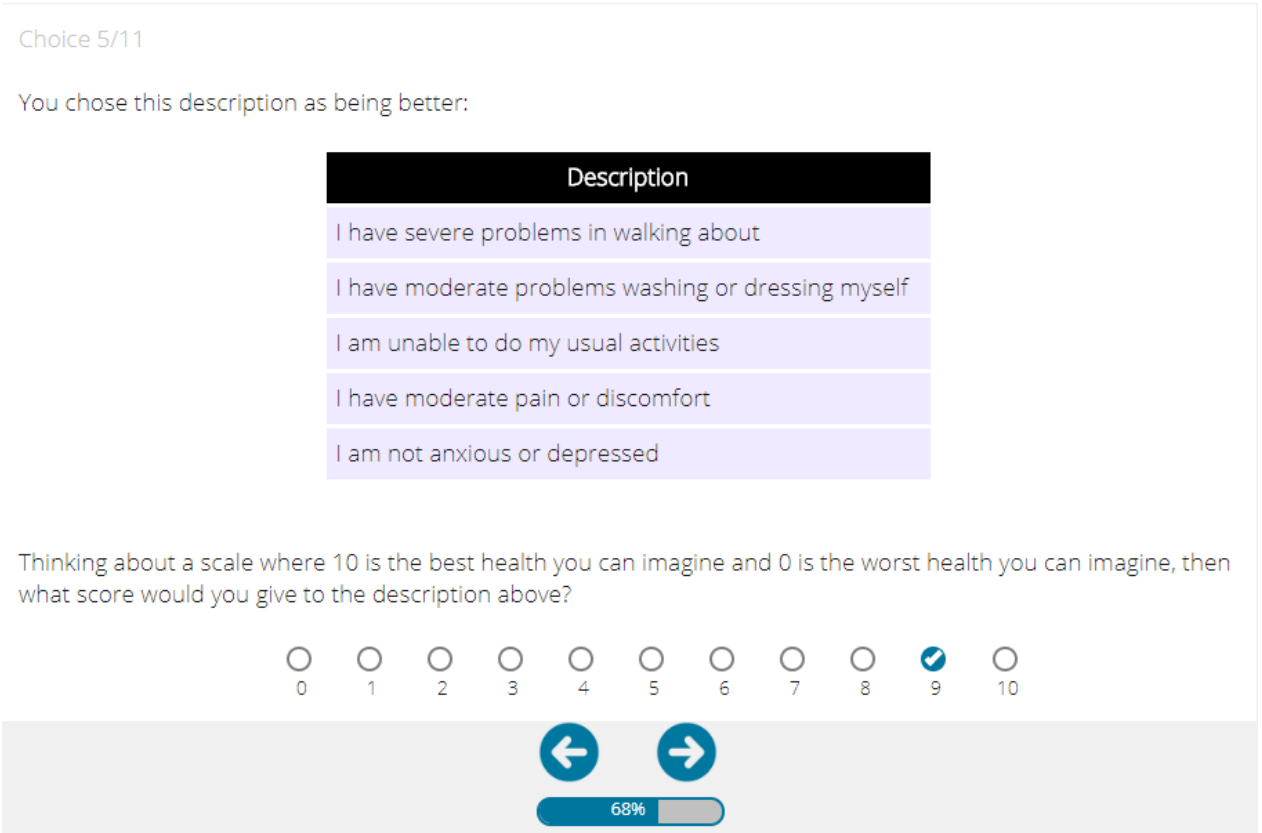


Figure 6: Example choice tasks from case study 3 discrete choice task

different from a level 2 parameter). For scale parameters, t-tests were used to assess if they were statistically different from 1. For each model, it was calculated whether the coefficient point estimates had the expected ordering ($0 > \beta_2 > \beta_3 > \beta_4 > \beta_5$). Model averaging was carried out with two/one-step models with and without scale parameters, and the posterior probability of participants being two/one-step decision-makers calculated from the results.

The models estimated using the main sample were used to predict for each choice task in the DCE the probability of observing (i) both choices and ratings, (ii) choices alone and (iii) ratings alone. The mean absolute difference between the predictions and observed frequencies of choices/ratings in the boost sample were then calculated for each model. All estimation was performed using the Apollo choice modelling package for R (Hess & Palma, 2019).

3.3.3 Results

The results of model estimation are given in Table 3. There were more statistically significant taste parameters in each of the COL models compared to MNL or OL. MNL and OL had eight and 12 statistically significant taste parameters respectively, whereas the two-step COL models had 19 without and 17 with a scale parameter, and one-step COL models had 18 without and 13 with a scale parameter.

The coefficient point estimates for usual activities levels 4 and 5 were illogically ordered in all models except the two one-step COL models, and in no case was the level 5 coefficient significantly lower than the level 4. In addition, the coefficient point estimates for self-care levels 4 and 5 in the OL model were illogically ordered, as well as the coefficients for self-care level 2 and mobility levels 2 and 3 in the MNL model.

Figure 7 illustrates the models' BIC. The combined MNL and OL models have a lower BIC than all COL models. For COL models, two-step models have a lower BIC than one-step, although differences are small.

Model averaging showed the percentage of two-step decision-makers was lower than in case study 2, at 57.5% and 70.0% respectively for models without and with a scale pa-

Table 3: Case study 3 model results

	MNL		OL		2-step no scale		1-step no scale		2-step scale		1-step scale	
Mobility L2	**-0.261**	(0.091)	-0.140*	(0.029)	-0.173*	(0.022)	-0.0851*	(0.025)	-0.251*	(0.028)	-0.126*	(0.038)
Mobility L3	**-0.161**	(0.088)	-0.323*	(0.026)	-0.261*	(0.019)	-0.231*	(0.022)	-0.316	(0.024)	-0.327*	(0.046)
Mobility L4	-1.14*	(0.095)	-0.674*	(0.037)	-0.776*	(0.023)	-0.773*	(0.026)	-0.989*	(0.03)	-1.11*	(0.12)
Mobility L5	-1.72*	(0.117)	-0.674	(0.04)	-1.01*	(0.027)	-1.08*	(0.03)	-1.30*	(0.036)	-1.56	(0.166)
Self-care L2	**0.113**	(0.085)	-0.225*	(0.026)	-0.162*	(0.018)	-0.250*	(0.022)	-0.170*	(0.023)	-0.354*	(0.047)
Self-care L3	-0.234*	(0.072)	-0.286	(0.026)	-0.270*	(0.017)	-0.418*	(0.021)	-0.306*	(0.022)	-0.588*	(0.063)
Self-care L4	-0.985*	(0.087)	**-0.688***	(0.032)	-0.703*	(0.021)	-0.856*	(0.024)	-0.879*	(0.026)	-1.23*	(0.13)
Self-care L5	-1.21	(0.088)	**-0.576**	(0.035)	-0.841*	(0.023)	-1.13*	(0.027)	-1.03*	(0.029)	-1.61	(0.167)
Usual activities L2	-0.121	(0.066)	-0.106*	(0.025)	-0.0984*	(0.016)	-0.160*	(0.02)	-0.103*	(0.021)	-0.228*	(0.037)
Usual activities L3	-0.157	(0.078)	-0.154	(0.026)	-0.152*	(0.017)	-0.194	(0.021)	-0.175	(0.022)	-0.277	(0.04)
Usual activities L4	**-1.04***	(0.079)	**-0.390***	(0.036)	**-0.603***	(0.022)	-0.761*	(0.025)	**-0.724***	(0.026)	-1.08*	(0.111)
Usual activities L5	**-0.901**	(0.079)	**-0.373**	(0.033)	**-0.584**	(0.021)	-0.763	(0.026)	**-0.709**	(0.026)	-1.09	(0.115)
Pain/discomfort L2	-0.304*	(0.077)	-0.166*	(0.029)	-0.130*	(0.02)	-0.148*	(0.024)	-0.173*	(0.025)	-0.218*	(0.043)
Pain/discomfort L3	-0.412	(0.085)	-0.320*	(0.031)	-0.304*	(0.02)	-0.358*	(0.026)	-0.376*	(0.025)	-0.518*	(0.066)
Pain/discomfort L4	-1.24*	(0.092)	-0.629*	(0.035)	-0.754*	(0.023)	-0.995*	(0.027)	-0.905*	(0.029)	-1.42*	(0.151)
Pain/discomfort L5	-1.48	(0.11)	-0.631	(0.04)	-0.957*	(0.027)	-1.22*	(0.032)	-1.18*	(0.034)	-1.75	(0.189)
Anxiety/depression L2	-0.221	(0.079)	-0.231*	(0.027)	-0.201*	(0.018)	-0.294*	(0.021)	-0.229*	(0.023)	-0.418*	(0.051)
Anxiety/depression L3	-0.512	(0.073)	-0.251	(0.028)	-0.308*	(0.018)	-0.392*	(0.022)	-0.372*	(0.023)	-0.563	(0.065)
Anxiety/depression L4	-1.52*	(0.103)	-0.760*	(0.037)	-0.881*	(0.026)	-1.12*	(0.03)	-1.08*	(0.032)	-1.61*	(0.17)
Anxiety/depression L5	-1.53	(0.091)	-0.763	(0.041)	-1.01*	(0.026)	-1.30*	(0.031)	-1.22*	(0.031)	-1.85	(0.193)
Scale parameter									0.583*	(0.14)	0.687*	(0.084)
τ_0			-6.51	(0.141)	-6.83	(0.139)	-5.03	(0.091)	-6.28	(0.073)	-4.97	(0.059)
τ_1			-4.82	(0.076)	-5.13	(0.069)	-4.02	(0.061)	-4.58	(0.055)	-3.97	(0.051)
τ_2			-3.58	(0.059)	-3.88	(0.05)	-3.17	(0.051)	-3.35	(0.048)	-3.13	(0.046)
τ_3			-2.58	(0.051)	-2.86	(0.042)	-2.4	(0.046)	-2.35	(0.043)	-2.36	(0.043)
τ_4			-1.8	(0.047)	-2.05	(0.038)	-1.73	(0.042)	-1.57	(0.04)	-1.69	(0.04)
τ_5			-0.945	(0.044)	-1.18	(0.035)	-0.95	(0.04)	-0.718	(0.038)	-0.916	(0.039)
τ_6			-0.143	(0.042)	-0.358	(0.035)	-0.181	(0.039)	0.0797	(0.04)	-0.149	(0.042)
τ_7			0.775	(0.044)	0.576	(0.039)	0.721	(0.042)	0.993	(0.06)	0.752	(0.062)
τ_8			1.93	(0.062)	1.74	(0.061)	1.87	(0.062)	2.14	(0.11)	1.9	(0.111)
τ_9			2.97	(0.11)	2.78	(0.111)	2.91	(0.111)	3.18	(0.018)	2.94	(0.073)
BIC	4773.9		140974.0		176992.5		177185.7		176328.8		177164.2	

Note. Standard errors in parentheses; * = significant at 5% level after adjustment using Holm's sequential Bonferonni correction; coefficients with illogical ordering in bold

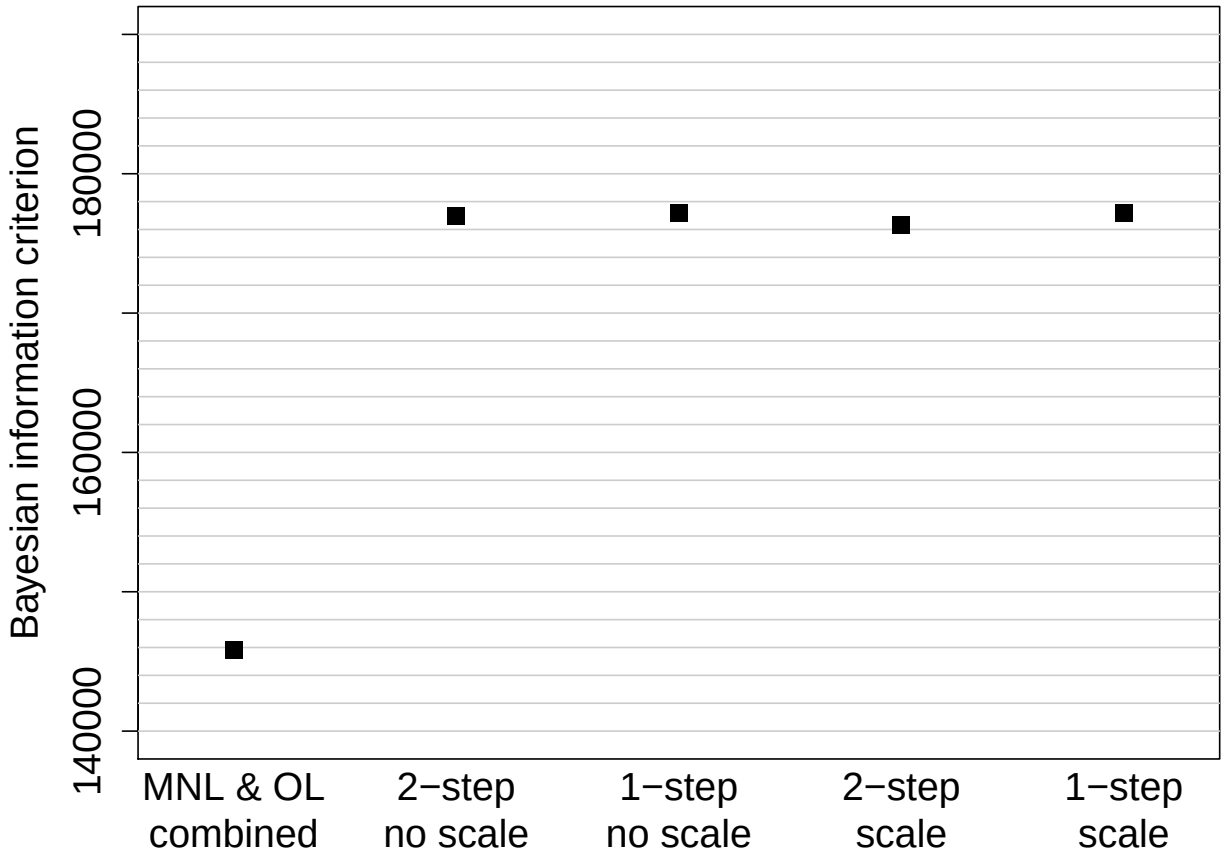


Figure 7: Bayesian information criteria for case study 3 models

Table 4: Comparison of absolute differences between out-of-sample forecasts and observed responses for case study 3

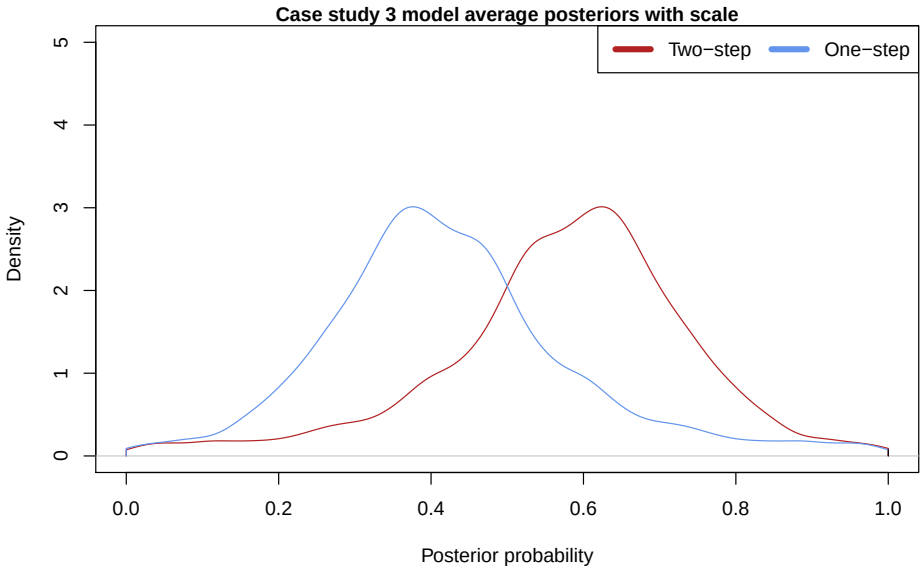
		MNL	OL	2-step no scale	1-step no scale	2-step scale	1-step scale
Choices and ratings	mean			0.0201	0.0345	0.0205	0.0204
	sd			0.0216	0.036	0.0222	0.0221
	95th percentile			0.0631	0.101	0.0665	0.0662
Choices only	mean	0.037		0.0944	0.286	0.0986	0.0954
	sd	0.0368		0.0571	0.146	0.061	0.0592
	95th percentile	0.1		0.188	0.535	0.2	0.193
Ratings only	mean		0.0307	0.038	0.0366	0.031	0.0307
	sd		0.0313	0.0428	0.0372	0.0317	0.0313
	95th percentile		0.0942	0.128	0.115	0.0955	0.0942

*Note.*sd = standard deviation

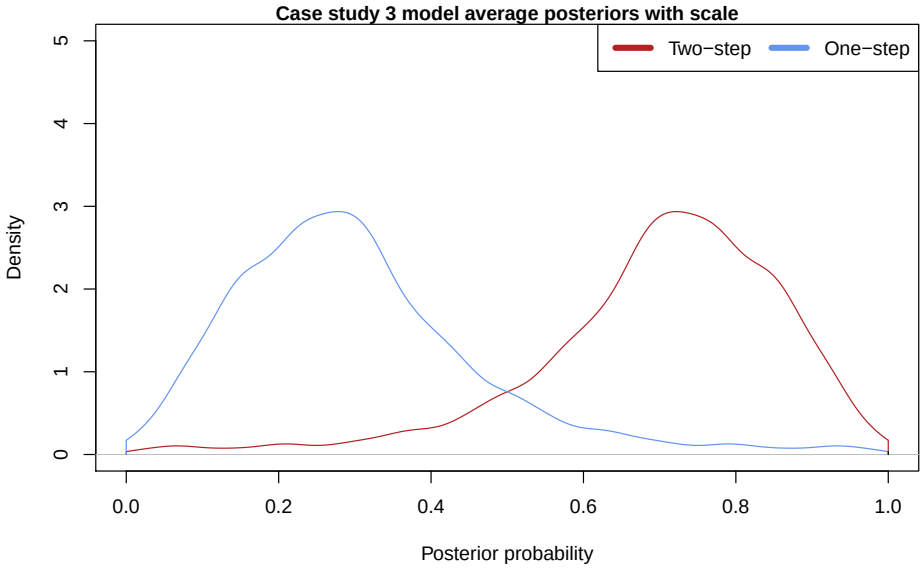
rameter. Figure 8 shows density plots of the posterior probabilities of belonging to each decision-making class. It can be seen that there is greater heterogeneity than in case 2 as the probability density is more spread out.

Figure 9 displays scatter plots of predicted choice-ordered probabilities from each model and the observed probabilities in the boost sample. For each task, the predicted and observed probabilities of each option being chosen and being given each rating. In each plot there are hence 10 tasks $\times$ 8 blocks $\times$ 2 options $\times$ 11 ratings = 1,760 data points. The patterns are similar for all models, with a tendency to over-predict low probabilities, and to under-predict high probabilities.

Table 4 summarises comparisons between out-of-sample predictions and observed responses. Differences between models are small, especially for the COL models when predicting joint choice-rating observations. Larger differences are seen when predicting choices alone, and it is notable that no COL model has a lower mean absolute difference between prediction and observation than the MNL model. The same is true when comparing OL to COL models in predicting ratings alone, however differences are small.

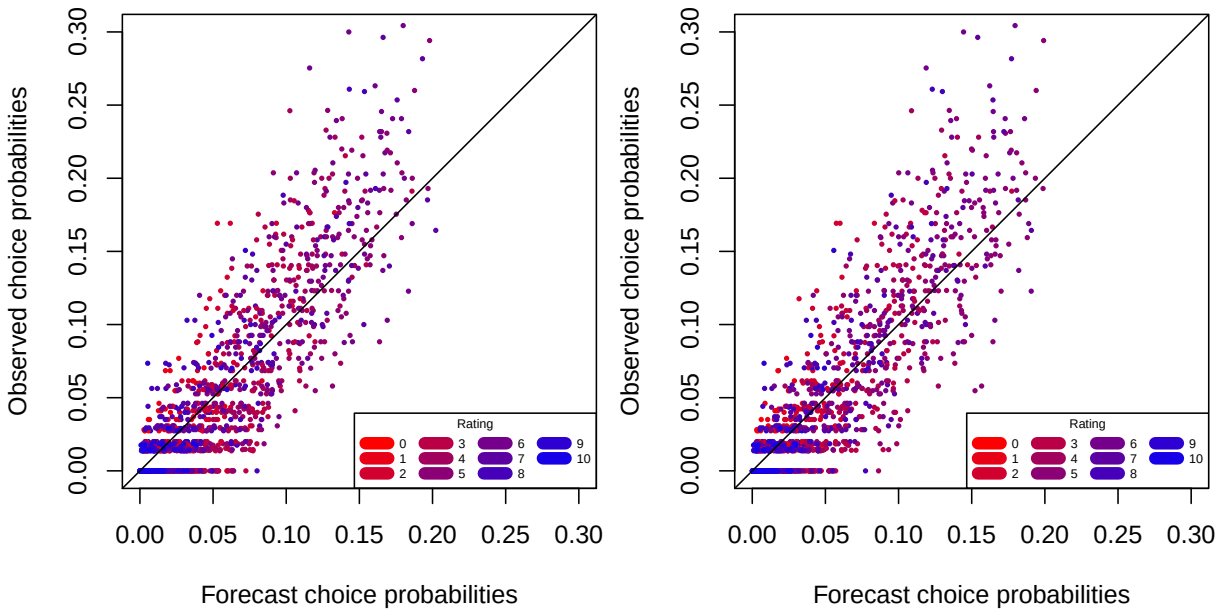


(a)



(b)

Figure 8: Model averaging posterior decision-making class allocation for case study 3



(a) 2 step

(b) 2 step scale

(c) 1 step

(d) 1 step scale

Figure 9: Scatter plots of out-of-sample forecasts/observed choices for choice-ordered models in case study 3

3.3.4 Discussion

As in case study 2, more coefficients achieved statistical significance with COL models compared to both MNL and OL, demonstrating a benefit to combining choice and rating data. The significance tests were also more rigorous than the standard test of being different from 0, as they assessed whether the coefficients followed a logical ordering. The logical ordering of coefficients is particularly important in this context, as to create a value set for a HRQoL instrument it is essential that the point estimates of the coefficients of worse states are lower than the point estimates of better states. By this metric, COL models also outperformed both MNL and OL: The point estimates of five coefficients in the MNL model and four coefficients in the OL model were illogically ordered. On the other hand, the point estimates of only two coefficients in the two-step COL models were illogically ordered, and all point estimates had the expected ordering for one-step COL models. (Although the differences between coefficients were not always statistically significant.)

In contrast to case study 2, here the combined MNL and OL models had a better model fit than any COL model. This may be due to the increased range of the Likert scale, with 11 possible ratings compared to only seven for case study 2. The extremes of the scale were chosen relatively rarely (states were rated 0 only 0.7% of the time, and rated 10 only 0.5% of the time), leading to models predicting ratings to have high log-likelihoods. It may be that using a separate MNL component that only had to predict a binary choice may result in the better fit when combined with OL compared to COL models which had to predict choices with the same coefficients that predicted ratings.

In this case study, unlike with case study 2, there is little reason why a positive probability should be assigned for choosing an option that is not also the highest rated. There are no plausible factors that should explain such a divergence between choices and ratings, thus a one-step decision-making process is arguably more appropriate. However, the two-step COL models have better fit than one-step models, albeit by a small amount.

More significantly, model averaging indicated that a majority of participants used a two-

step decision-making process. Yet comparing the results to case study 2, a greater fraction of one-step decision-makers were observed, and there was greater heterogeneity in the posterior probabilities of class membership. Combined with the large sample size, there is evidence that many participants were using a one-step decision-making process.

This case study illustrates that COL models can predict out-of-sample responses with some success. Even though a standard MNL model is superior on average at predicting choices alone, and OL models were superior for ratings alone, it is not by much.

Many studies approach out-of-sample forecasting by randomly dividing a single dataset in two, thus the estimation and prediction data have identical characteristics (at least in expectation). In this case study, the main sample used for estimation and the boost sample used for prediction differ, especially in terms of age and gender. This approach has some advantages, in that it is a more robust test of a model's ability to predict out-of-sample. On the other hand, a disadvantage is that the different target populations for the main and boost samples makes the patterns of deviations between predictions and observations difficult to interpret.

4 General discussion

In the above demonstrations of applying COL models, there was not necessarily an advantage over standard MNL and OL models. Thus in cases where the researcher is only interested in either choices or ratings, there is a case for using existing models such as MNL or OL.

However, if both choices and ratings are of interest, and the observations are created in a joint process, it seems logical to use a modelling approach that takes account of the joint nature of the data. Even if only choices are primarily of interest, there may be benefits to using a joint approach, as was demonstrated in case study 3. There, the research question was the relative valuation of EQ-5D levels, which may be obtained using discrete choices alone. Yet only when using COL models did all coefficients have the expected signs, as

required for constructing a quality of life value set. In addition, if individuals are using a one-step decision-making process with choices and ratings, then both MNL and OL (and more complicated versions of them such as mixed logit) are misspecified.

Although joint modelling of choices and ratings may not necessarily give an advantage as measured by performance metrics, adding ratings to choices does have a conceptual advantage in that it allows measurement of strength of preference. This benefit is most readily seen in case 2. The research question addressed participants' preferences for different AAC systems, and how these change if choosing for different children. Analysis of discrete choices alone allows the modeller to create lists of all possible systems in order of preferences for each child. However, it is only when combined with rating data that the strength of preference can be assessed. Thus, for example, participants' ranking of systems could be identical for two children, yet for one child a given system is much more preferred than all others, and for the other child the preference is only weak.

It was argued that a two-step decision-making process was reasonable to assume for case study 2, whereas a one-step model was reasonable to assume for case study 3. In practice, two-step models provided better model fit than one-step models with both datasets, and model averaging revealed that the majority of individuals' choices were more consistent with a two-step decision-making process. Yet in line with the arguments above, the fraction of one-step decision-makers was higher in case study 3 than in case study 2, and case study 3's one-step models were the only ones to have all coefficients logically ordered. In fact, it could be argued that in case study 3, one step decision-making should be a normative standard, and two-step represents a deviation from logical decision-making.

Identifying whether individuals were using one- or two-step decision-making relied on assuming that individuals interpreted the rating scale in the same way throughout. There are several reasons why this assumption may not have held, for example there could have been learning effects at the start of surveys and/or fatigue towards the end. In addition, respondents may have re-calibrated their internal scale as the surveys progressed and they

were exposed to more items. Such effects could lead to one-step decision-making to appear more like two-step. In case study 2, this did not seem to present any significant issues, as participants were expected to use two-step decision-making, which was borne out by model averaging estimating that around 80% did so. The results of one-step model estimation also appeared similar to case study 1 when misspecified one-step models were estimated on a true two-step data generating process. However, in case study 3 model averaging estimated that 30-40% of participants used two-step decision-making, despite one-step being anticipated. This relatively large fraction indicates that some one-step decision-makers may potentially have been incorrectly specified as two-step. Future research could investigate learning, fatigue or rating scale adjustment, for example by introducing individual level heteroscedasticity or by assessing response consistency using identical first and last tasks.

Case studies 2 and 3 have illustrated that whether to use a two- or one-step model in future studies is a complex issue. There are measures of model performance, such as BIC, which could be used to guide modelling choices. However, there may also be important theoretical considerations. It was argued in case study 3 that one-step decision-making ought to be a normative standard. That was the case as an underlying theoretical assumption of valuing instruments like EQ-5D is that individuals position all health states on a univariate value scale, so that they should never choose a state which is not also the highest ranked. With case study 2, on the other hand, it was expected that individuals would use two-step decision-making, but this was not considered a normative standard. For case study 2 tasks, there were many plausible reasons why participants might not choose the highest rated alternative. However, there was nothing illogical in always choosing the highest rated alternative, and one-step decision-making would not violate any theoretical assumptions. Whether to use a one- or two-step model must be decided on a case-by-case basis, and guidance for future studies is to carefully assess both theoretical implications and empirical measures of model performance in doing so.

Whether or not to include a scale parameter is easier to address, since there are fewer

theoretical implications of including or not including one. Guidance for future studies is to use conventional methods to decide whether to include a scale parameter, for example t-tests of parameter significance, or using likelihood ratio tests or an information criterion to choose between alternate models.

It is a strength of the current paper that it has illustrated the application of COL models in several different situations with different sources of data. The three case studies have also each highlighted different aspects of using COL models. The data was either synthetic or was stated preference data from hypothetical surveys. Using stated preference data had the advantage that individuals' choice sets were well defined. However, a disadvantage of this study is that it does not examine how well COL models would perform with revealed preference data.

Future research could usefully study how to extend COL models for use with revealed preference data, for example online retail data, with consumers selecting items, then rating them. Another possibility in health may be to study patients who have a choice of treatment centres, and who subsequently provide patient satisfaction ratings. A potential hurdle to be overcome in each case would be to define individuals' consideration sets. In addition, with revealed preference data there will typically be another decision-making step not included in the current formulation of COL models. Individuals first make a choice, then decide whether to rate, then decide what rating to give. The first and third steps are captured by COL models, but the second is not, and future research could explore how to incorporate it into the modelling framework.

The case studies in this manuscript all involved individuals making repeated choices and ratings. This is typical for stated preference exercises, and also some revealed preference data, for example watching and rating films and TV episodes on a streaming service. Yet there are many situations in which individuals make one-off decisions, with potentially quite different properties to repeated choice. Future research is needed to assess the usefulness of COL models when individuals make one-off decisions.

Although some evidence of the ability of COL models to provide out-of-sample forecasts was shown in case study 3, it would be useful to examine their performance if the estimation and evaluation samples differ more than they did here.

5 Conclusion

The usefulness of COL models has been demonstrated, and they should be considered when modelling data with both discrete choices and ratings. In addition, researchers carrying out discrete choice experiments should consider adding to their survey a Likert scale question rating their chosen alternative. This allows strength of preference to be assessed, and as only the chosen alternative is rated, it minimises the additional burden on participants.

There is much future work to be carried out on COL models. In particular, participant heterogeneity could be explored by introducing random parameters. Another possibility is allowing some parameters to affect only choices or ratings.

Acknowledgements

We would like to thank the following people who were instrumental in gathering the data used in case study 2: David Meads (University of Leeds), Yvonne Lynch, Juliet Goldbart, Stuart Meredith, Liz Moulam, Janice Murray (Manchester Metropolitan University), Nicola Randall and Simon Judge (Barnsley Hospital NHS Foundation Trust).

We would also like to thank Paul Kind and Francesca Torelli (University of Leeds), who were instrumental in gathering the data used in case study 3.

Data collection for case study 2 was funded by the NIHR Health Services and Delivery Program (project 14/70/153). The views expressed are those of the authors, and not necessarily those of the NHS, the NIHR, or the Department of Health.

Data collection for case study 3 was funded by Prostate Cancer UK.

Stephane Hess acknowledges additional support by the European Research Council through

the consolidator grant 615596-DECISIONS.

References

Bansback, N., Brazier, J., Tsuchiya, A., & Anis, A. (2012). Using a discrete choice experiment to estimate health state utility values. *Journal of health economics*, *31*(1), 306–318.

Beck, M. J., Fifer, S., & Rose, J. M. (2016). Can you ever be certain? reducing hypothetical bias in stated choice experiments via respondent reported choice certainty. *Transportation Research Part B: Methodological*, *89*, 149–167.

Beck, M. J., Rose, J. M., & Hensher, D. A. (2013). Consistently inconsistent: The role of certainty, acceptability and scale in choice. *Transportation Research Part E: Logistics and Transportation Review*, *56*, 81–93.

Devlin, N. J., Shah, K. K., Feng, Y., Mulhern, B., & van Hout, B. (2018). Valuing health-related quality of life: An eq-5 d-5 l value set for e ngland. *Health Economics*, *27*(1), 7–22.

Dolan, P. (1997). Modeling valuations for euroqol health states. *Medical care*, 1095–1108.

Gutknecht, M., Schaarschmidt, M.-L., Danner, M., Blome, C., & Augustin, M. (2018). Measuring the importance of health domains in psoriasis–discrete choice experiment versus rating scales. *Patient preference and adherence*, *12*, 363.

Hancock, T. O., & Hess, S. (2020). *What is really uncovered by mixing different model structures: contrasts between latent class and model averaging.* Retrieved from `https://www.stephanehess.me.uk/papers/working%20papers/Hancock_Hess_2020_model_averaging.pdf`

Hensher, D. A., & Rose, J. M. (2012). The influence of alternative acceptability, attribute thresholds and choice response certainty on automobile purchase preferences. *Journal of Transport Economics and Policy (JTEP)*, *46*(3), 451–468.

Hess, S. (2014). Latent class structures: taste heterogeneity and beyond. In S. Hess &

A. Daly (Eds.), *Handbook of choice modelling* (p. 311-332). Edward Elgar Publishing.

Hess, S., & Daly, A. (2014). *Handbook of choice modelling*. Edward Elgar Publishing.

Hess, S., & Palma, D. (2019). Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of Choice Modelling*, 100170.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Mattmann, M., Logar, I., & Brouwer, R. (2019). Choice certainty, consistency, and monotonicity in discrete choice experiments. *Journal of Environmental Economics and Policy*, *8*(2), 109–127.

Moe, W. W., & Schweidel, D. A. (2012). Online product opinions: Incidence, evaluation, and evolution. *Marketing Science*, *31*(3), 372–386.

Moorcroft, A., Scarinci, N., & Meyer, C. (2019). A systematic review of the barriers and facilitators to the provision and use of low-tech and unaided aac systems for people with complex communication needs and their families. *Disability and Rehabilitation: Assistive Technology*, *14*(7), 710–731.

Mulhern, B., Bansback, N., Hole, A. R., & Tsuchiya, A. (2017). Using discrete choice experiments with duration to model eq-5d-5l health state preferences: testing experimental design strategies. *Medical Decision Making*, *37*(3), 285–297.

Murray, J., & Goldbart, J. (2009). Augmentative and alternative communication: a review of current issues. *Paediatrics and child health*, *19*(10), 464–468.

Ramos-Goñi, J. M., Pinto-Prades, J. L., Oppe, M., Cabasés, J. M., Serrano-Aguilar, P., & Rivero-Arias, O. (2017). Valuation and modeling of eq-5d-5l health states using a hybrid approach. *Medical care*, *55*(7), e51–e58.

Regier, D. A., Watson, V., Burnett, H., & Ungar, W. J. (2014). Task complexity and response certainty in discrete choice experiments: an application to drug treatments for juvenile idiopathic arthritis. *Journal of Behavioral and Experimental Economics*, *50*, 40–49.

Rose, J. M., Beck, M. J., & Hensher, D. A. (2015). The joint estimation of respondent-

reported certainty and acceptability with choice. *Transportation Research Part A: Policy and Practice*, *71*, 141–152.

Soekhai, V., de Bekker-Grob, E. W., Ellis, A. R., & Vass, C. M. (2019). Discrete choice experiments in health economics: past, present and future. *PharmacoEconomics*, *37*(2), 201–226.

Stolk, E. A., Oppe, M., Scalone, L., & Krabbe, P. F. (2010). Discrete choice modeling for the quantification of health states: the case of the eq-5d. *Value in Health*, *13*(8), 1005–1013.

Sun, M. (2012). How does the variance of product ratings matter? *Management Science*, *58*(4), 696–707.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Webb, E. J., Lynch, Y., Meads, D., Judge, S., Randall, N., Goldbart, J., ... Murray, J. (2019). Finding the best fit: Examining the decision-making of augmentative and alternative communication professionals in the UK using a discrete choice experiment. *BMJ open*, *9*(11).

Webb, E. J., Meads, D., Lynch, Y., Randall, N., Judge, S., Goldbart, J., ... Murray, J. (2019). What's important in AAC decision making for children? Evidence from a best–worst scaling survey. *Augmentative and Alternative Communication*, 1–15.

Wijnen, B. F., van der Putten, I. M., Groothuis, S., de Kinderen, R. J., Noben, C. Y., Paulus, A. T., ... Hiligsmann, M. (2015). Discrete-choice experiments versus rating scale exercises to evaluate the importance of attributes. *Expert review of pharmacoeconomics & outcomes research*, *15*(4), 721–728.

Appendix

Table A.1: Case study 2 attributes and levels

Child attribute	Levels
Receptive and expressive language	Delayed
	Receptive language exceeding expressive language
Communication ability with AAC	No previous AAC experience
	Able to use AAC for a few communicative functions
	Able to use AAC for a range of communicative functions
Child's determination and persistence	Does not appear motivated to communicate through any methods and means
	Motivated to communicate through symbol communication systems
	Only motivated to communicate through methods other than symbol communication
Predicted future skills and abilities	Regression
	Plateau
	Progression
AAC system attribute	**Levels**
Vocabulary sets	No vocabulary set
	Fixed vocabulary set
	Vocabulary set with staged progression
Consistency of layout	Consistency of some aspects of layout
	Consistency of all aspects of layout
	Idiosyncratic layout
Type of vocabulary organisation	Visual scene
	Taxonomic
	Semantic-syntactic
	Pragmatic
Size of vocabulary	Up to 50 vocabulary items
	50-1000 vocabulary items
	More than 1000 vocabulary items
Graphic representation	Photos
	Pictographic symbol set
	Ideographic symbol system (with rules or encoding)
	Text