How (not) to evaluate respondents' data quality in the upcoming Discrete Choice Experiment Valuation Technology (EQ-DCE-VT) protocol

Dr. Marcel F. Jonker[*] [1,2]

1. Erasmus School of Health Policy & Management, Erasmus University Rotterdam, The Netherlands
2. Erasmus Choice Modelling Centre, Erasmus University Rotterdam, The Netherlands

* corresponding author:

Dr. Marcel F. Jonker
Erasmus University Rotterdam
PO Box 1738
3000DR Rotterdam
The Netherlands
Email: marcel@mfjonker.com

ABSTRACT

OBJECTIVES: The upcoming stand-alone discrete choice experiment (DCE) valuation protocol requires the Group to think about efficient and reliable approaches to assess DCE data quality. This manuscript introduces the garbage class mixed logit (MIXL) model as a convenient and performant alternative to manually screening for respondents with low data quality.

METHODS: Garbage classes are typically used in latent class logit analyses to designate or identify group(s) of respondents with low data quality. Yet the same concept can be applied to achieve an automated selection of respondents in MIXL models as well.

RESULTS: Based on a re-analysis of four DCEs, including an EQ-5D-5L dataset, it is shown that the garbage class MIXL model and root likelihood (RLH) tests have indistinguishable empirical accuracy. Previous research has shown that the latter has superior performance compared to internal validity tests (such as repeated and dominant choice tasks), which means that also garbage class MIXL models have excellent sensitivity and specificity. The advantage of garbage class MIXL models, however, is that they require no user effort and produce preference estimates that do not depend on statistical cut-off values.

CONCLUSIONS: Including a garbage class in MIXL models removes the influence of respondents with a random choice pattern from the MIXL model estimates, provides an estimate of the number of low-quality respondents in the dataset, and avoids having to manually screen for respondents with low data quality based on internal and/or statistical validity tests. Although less versatile than the combination of standard MIXL estimates with separate assessments of data quality and sensitivity analyses, the proposed garbage class MIXL model provides a fully automated and reliable alternative that is applicable to both DCE with and without duration data but particularly relevant for the upcoming EQ-DCE-VT protocol.

INTRODUCTION

The Valuation Working Group (VWG) and EQ-funded researchers have for many years been working on a stand-alone discrete choice experiment (DCE) with duration methods for health-state valuations. At the EuroQol Academy meeting in Prague (2020), evidence with respect to the layout and design of choice tasks, optimization of experimental designs, and modelling strategies was presented and discussed in an open session. This resulted in a draft protocol[1] that is currently being piloted in Trinidad and Tobago. Final results are expected before the end of the year; however, based on a very similar protocol, the wellbeing of older people (WOOP) instrument has recently been valued without encountering any problems,[2] which seems to confirm that the protocol works well and is nearing its official introduction.

Value sets for the adult versions of the EQ-5D (i.e. EQ-5D-3L and EQ-5D-5L instruments) are currently still based on composite TTO (cTTO) data collected using face to face interviews, generally supplemented with DCE data that are collected during the same interview. Value sets for the pediatric version of the EQ-5D (i.e. EQ-5D-Y-3L instrument) are based on online DCE surveys that are supplemented with a small cTTO sample that allows the estimated DCE values to be re-scaled to the QALY scale.

To ensure adequate data quality in the TTO-part of valuation studies, the EuroQol Group has created a valuation protocol (EQ-VT) that includes stringent checks on interviewer quality and consistency in the cTTO tasks. For example, response data are considered of insufficient quality when the interviewer does not explain all aspects of the task during the warmup (e.g. when the respondent does not enter the lead-time TTO task during the example questions), spends an unreasonably small amount of time on the practice tasks, spends an unreasonably small amount of time on the valuation tasks themselves, or exhibits severely inconsistent TTO values).

In contrast, checks on the DCE response data are much more limited and focus on the identification of suspicious response patterns, such as flatlining, and identification of respondents with unusually fast DCE completion times, so-called ‘speeders’. For the EQ-5D-Y-3L DCE data there is no official quality control protocol (yet) but thus far dominant pairs and DCE completion times have been used as quality checks in all published [3-6] and recently submitted studies.

One of the reasons why the VWG has not yet created an official quality control protocol for DCE data is that there are no widely accepted or "golden" indicators of response quality in DCEs. Internal validity tests, which are checks on the logic, consistency, and trade-off assumptions in the discrete choice data, have been recommended to screen for respondents with low response quality.[7-9] The dominant choice tasks that have thus far been used in the EQ-5D-Y studies are examples of such tests. The problem with internal validity tests, however, is that they cannot take response error into account and are consequently inconsistent with the underlying theoretical framework of DCEs. As a result, violations of internal validity tests are notoriously difficult to interpret, particularly when the predicted utility difference between the included choice options is small.[10-11] Moreover, in recent years there has been a clear shift from conditional logit models to statistical models that accommodate preference heterogeneity, with the mixed logit (MIXL) model currently being the most commonly used model to analyze discrete choice data[12] and also being the type of model selected in the draft stand-alone DCE valuation protocol.[1] In these models the predicted utility difference between identical choice options varies between respondents, which makes it even more challenging to correctly interpret violations of internal validity tests.[11] More importantly, also from an empirical perspective the performance of internal validity tests has been shown to be inadequate. For example, as explained in last year's Plenary paper on the sensitivity and specificity of repeated and dominant choice tasks, the predictive accuracy of repeated and dominant choice tasks is only slightly better than a random coin flip. In contrast, likelihood-based statistical validity tests, such as the root likelihood (RLH) statistic, can provide a superior alternative for the assessment of respondents' response quality in DCEs.[11]

The use of the RLH statistic, however, also has some inherent limitations. Most importantly, the use of the RLH statistic involves a laborious process that requires practitioners to fit a standard MIXL, compute respondent-specific RLH statistics and associated uncertainty measures, and then select one or more statistical cut-off values to be able to classify respondents as either having provided good or bad quality responses. Subsequently, assuming the ultimate goal is to present statistical estimates that are unaffected or shown to only be marginally affected by respondents with low-quality response patterns, those identified as bad-quality respondents need to be excluded from the sample and additional models need to be fit. In addition to the required effort and estimation time, the subjective selection of one or more statistical cut-off values thus introduces the possibility for practitioners to manipulate the reported preference estimates–either deliberately or unintentionally. This makes the RLH statistic an interesting and valuable approach to assess respondents' response quality in DCEs, but also an approach that requires substantial effort and introduces a certain degree of ambiguity in published estimation results.

In this paper, a different approach to accounting for respondents with low data quality in DCEs is proposed: one that is based on a latent-class MIXL model with two classes. The first class represents the standard MIXL model that one would normally fit (e.g. when computing RLH statistics), whereas the second class represents a so-called 'garbage class' in which respondents make arbitrary choices between the choice options in each task. The inclusion of a garbage class has substantial similarities with scale-adjusted latent class (SALC) logit models as introduced by Magidson and Vermunt[13], particularly those in which one of the scale classes has a scale constrained to zero.[14] However, to the best of my knowledge, garbage classes have thus far not been combined with a MIXL model specification.

The intuition of the proposed garbage class MIXL model is very similar to how a standard latent class logit model works. During model estimation, each respondent is assigned with a certain probability to the standard MIXL specification and with one minus that probability to the garbage class; this is based on the match between the respondents' response patterns and the two utility functions. The estimated class-membership probability indicator thereby provides an easy-to-interpret alternative to the RLH statistic: at the individual-level, the class-selection probability can be used to detect respondents with a response pattern that better fits the garbage class than the standard MIXL model, whereas at the population level the aggregate class-membership probability provides an indication of the number of respondents with a low-quality response pattern in the dataset. Also, by including a garbage class in the MIXL specification, the preference estimates automatically only reflect the preferences of the good-quality respondents, i.e. without having to manually conduct split sample analyses based on internal or statistical validity tests. Hence the MIXL potentially provides a convenient alternative that is also directly applicable to the type of models that need to fitted when tariffs need to be presented.

In the remainder of this paper, the garbage class MIXL model is first formally introduced and subsequently compared with that of a standard MIXL model with respondent-selection based on RLH statistics. Based on the presented similarity between both approaches in four different datasets that were previously analyzed using standard MIXL models, including one based on the EQ-5D-5L, the MIXL model with a garbage class is presented as a simpler alternative to manually screening for respondents with low data quality. Instead, it automatically provides MIXL estimates that are unaffected by respondents with low-quality response patterns in addition to estimates of the number of low-quality respondents in the dataset and, if required, classification of individual-level respondents into good/bad quality participants.

METHODS

Standard MIXL model with RLH statistics

In a MIXL model it is typically assumed that there are N respondents that each complete T discrete choice tasks, each consisting of J alternatives that are described by K explanatory variables. All explanatory variables can then be summarized in the design matrix

$$X_{itjk}\ (\mathbb{R}) \text{ for } \mathrm{i} = 1, \dots, \mathrm{N};\ \mathrm{t} = 1, \dots, \mathrm{T};\ \mathrm{j} = 1, \dots, \mathrm{J};\ \mathrm{k} = 1, \dots, \mathrm{K} \qquad (1)$$

and all observed choices in the response vector

$$Y_{itj} \in \{0,1\}\,, \qquad (2)$$

with the dependent variable (Y) being equal to 1 for the alternative that was chosen and zero for all other alternatives in each choice task, i.e.

$$(Y_{itj} = 1) \implies (Y_{itm} = 0\,, \forall m \neq j)\,. \qquad (3)$$

Following Random Utility Theory (RUT), each respondent is presumed to have chosen the option that provides them the highest utility

$$U_{itj} = V_{itj} + \epsilon_{itj} \qquad (4a)$$

with V denoting the structural (logical) part of the utility function and ϵ the error term.

A. The structural part of the utility function is typically defined as a linear additive function

$$V_{itj} = \sum_{k=1}^{K} \beta_{ik} * X_{itjk} \qquad (5)$$

with β_i denoting a vector of K coefficients that can take any desired joint distribution $f(\beta|\theta)$ across respondents and with θ denoting the coefficients of the joint distribution.

B. The error term is assumed to be independently and identically Gumbel distributed. Accordingly, the probability of choosing alternative j in choice task t is defined as

$$P_{itj} = \frac{\exp(V_{itj})}{\sum_{k=1}^{J}\exp(V_{itk})} \quad . \tag{6}$$

In the standard MIXL model, the mixing distribution of the respondents' β-coefficients is assumed to be multivariate normal (MVN) with mean vector μ and covariance matrix Σ, that is

$$f(\beta) \sim MVN(\mu, \Sigma). \tag{7}$$

Although different distributions can be specified without loss of generality, the MIXL models in this paper will make the same assumption. Based on the chosen MVN mixing distribution, the likelihood contribution of each individual respondent is given by:

$$L_i = (2\pi)^{-\frac{1}{2}V} \, |\det(\Sigma)|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\langle(\beta_i - \mu)' \, \Sigma^{-1} \, (\beta_i - \mu)\rangle + \right.$$
$$\left. \sum_{t=1}^{T}\sum_{j=1}^{J}(Y_{itj} \log(P_{itj}))\right] d\beta \tag{8}$$

and the individual-level root likelihood (RLH) statistic is defined as the geometric mean of the respondents' likelihood across the T choice tasks in the DCE

$$RLH_i = (L_i)^{1/T} \quad . \tag{9}$$

Interestingly, a null model with equal choice probabilities for all choice options (i.e. based on entirely random choice patterns) results in a RLH of 1/J. As such, respondents with $\text{RLH}_\text{i} \leq 1/J$ could be defined as 'bad quality' respondents. However, to appropriately take the statistical uncertainty of the RLH estimate into account, it makes sense to compute the probability that the RLH estimates are less than or equal to 1/J and to compare this statistic with different cut-off values (e.g. 0.01, 0.05 and 0.10).[11] If the estimated probability is larger than the chosen cut-off value, respondents are classified as having provided low-quality responses, or as (sufficiently) good-quality responses otherwise.

MIXL model with a garbage class

The garbage class MIXL model represents a relatively simple extension of the standard MIXL model in the sense that the only required adjustment of the model specification is the multiplication of the structural component of the utility function (V_{itj}) with a class membership parameter (φ_i)

$$U_{itj} = \varphi_i * V_{itj} + \epsilon_{itj} \ . \tag{4b}$$

No other changes are necessary and the interpretation of the class-membership parameter is also straight-forward: φ_i represents the respondent-specific probability of being assigned to the standard MIXL utility specification; if φ_i equals 1, respondents are entirely assigned to the standard MIXL utility specification ($U_{itj} = V_{itj} + \epsilon_{itj}$), and if φ_i equals 0 there is no contribution from the structural part of the utility function and respondents make choices entirely based on the error term ($U_{itj} = \epsilon_{itj}$). When aggregated, the sample average φ_i represents the class share of the MIXL model and the sample average of $1 - \varphi_i$ the garbage class share, which is of particular interest because it provides a readily available estimate of the number of low-quality respondents in the dataset. Similar to standard latent class logit models, it often makes sense to model the class membership using a binary logit model:

$$\varphi_i = \frac{\exp(\gamma \mathrm{Z}_i)}{1+\exp(\gamma \mathrm{Z}_i)} \tag{10}$$

where Z_i and γ denote the set of included class-membership predictor variables and corresponding vector of class membership model parameters, respectively. However, to appropriately compare the garbage class MIXL model with the RLH method, which does not rely on covariates, the φ_i parameters in this paper are estimated directly.

In a garbage class MIXL model, the estimated preference parameters have the advantage to only reflect the preferences of the good-quality respondents. This is different from the standard MIXL model, in which the preference parameters need to simultaneously reflect the choice patterns of the good and bad quality respondents. Accordingly, both models produce identical preference estimates if there are no low-quality respondents identified in the dataset. However, the more respondents with a low-quality response pattern, the more the standard MIXL estimates would be biased towards zero with a corresponding increase in the relative size of standard deviations of the mixing distribution. In contrast, the estimates of the MIXL model with a garbage class would remain unaffected, without the need to manually select respondents based on arbitrary statistical cut-off values and without having to produce sensitivity analyses based on the manual exclusion of low-quality respondents.

Model estimation

Both the standard and the MIXL model with a garbage class can be estimated using simulated maximum likelihood methods (e.g. using Apollo[15] or Biogeme[16] software) but can also be conveniently estimated using Bayesian Markov-Chain Monte Carlo (MCMC) methods. The latter involves the selection of prior densities for the model parameters and updating these based on the likelihood of the data, which is the approach that is used in this paper. Uninformative multivariate normal priors (i.e., with a mean of zero and standard deviation of 10) were assigned to μ, Bernoulli(0.5) priors to the φ_i parameters, and a Wishart prior with an identity scale matrix and K degrees of freedom to the inverse variance-covariance matrix (i.e. Σ^{-1}). Accordingly, the MIXL specification allows for potentially correlated preference parameters. Standard Gibbs update steps were used to update μ and Σ^{-1}, slice update steps to update φ_i, and a Metropolis-within-Gibbs algorithm with antithetic sampling as described by Bédard et al. (2014) to update the β_i parameters. All estimations were performed using the OpenBUGS software package[17] and were based on 25,000 MCMC iterations to let 3 MCMC chains converge and 75,000 iterations to reliably approximate the posterior distribution. Convergence was evaluated based on a visual inspection of the chains and the convergence diagnostics as implemented in the OpenBUGS package.

Datasets

The performance of the garbage class MIXL model was compared with that of the standard MIXL model combined with RLH estimates based on a re-analysis of four health-related DCEs. In a previous Plenary paper, these DCEs were used to assess the sensitivity and specificity of repeated and dominant choice tasks in DCEs in comparison with that of the RLH statistic.[18] Table 1 provides an overview of the DCE topics, the type of DCE designs that were used, and the DCE and dataset dimensions. Briefly summarized, all four datasets were collected using DCE instruments that were a replication of previously existing publications. Hence the attributes, levels and visual layouts were already tested and verified by the original authors. All data were collected via unattended online MTurk surveys. This ensured a mixture of good and bad quality respondents, which is essential for a meaningful comparison between the standard and garbage class MIXL model. In each DCE, both the order of the choice tasks and position of the choice options per choice task were randomized. All respondents received a small financial compensation for successfully completing the survey and, to ensure approximately US nationally representative samples, stratified quota sampling was implemented based on sex (male/female) and age groups (18-34/35-54/55-74/75+). Appendix C in the Online Supplemental found at https://doi.org/10.1016/j.jval.2022.01.015 provides a detailed overview of the sample representativeness and survey drop-out rates in each of the four datasets.

Table 1. Overview of the four datasets that are used in the MIXL model comparisons

DCE topic *	DCE design	# attributes/levels	optout	# parameters	# respondents	# choice tasks per respondent	# choice options per choice task
1. Antibiotics [7]	Full factorial **	2/2/2	No	3	750	13	2
2. Vaccines [8]	Bayesian D-efficient ***	4/3/3/3/2	Yes	11	500	16	3
3. Meals [9]	Bayesian D-Efficient ***	3/3/4/4/4	Yes	14	500	14	5
4. EQ-5D-5L [10]	Bayesian D-Efficient ***	5/5/5/5/5	No	20	500	21	2

* Citations of the original publications in parentheses. ** Each respondent completed all possible (non-dominant) pairwise choice tasks
*** Respondents completed one subdesign of a heterogeneous Bayesian DCE design with 10 subdesigns [25]

Comparison #1. Similarity between respondent classification at the sample level

The first comparison between the standard and garbage class MIXL model looked at the percentage of respondents that are classified as having a good or bad quality response pattern in each of the four datasets. For the standard MIXL model, based on Jonker et al.[1], three increasingly more conservative cut-off values were specified based on the mean posterior probability of the respondents' RLH<1/J (i.e. low-response quality) being larger than 0.01, 0.05 and 0.10, respectively. For the garbage class MIXL model, similar cut-off values were specified based on the mean posterior probability of φ_i<0.50 (i.e. garbage class membership) being larger than 0.20, 0.75, and 0.95, respectively. These latter values were chosen to approximately maximize the similarity between both methods across the four datasets. Obviously, the more similar both approaches are, the smaller the achieved minimum absolute differences between the methods will be.

Comparison #2. Similarity between individual-level respondent classifications

The second model comparison was based on the total number of respondents that are identically classified by both approaches. More specifically, for each of the three increasingly more conservative cut-off values, the percentage of respondents that are identically classified by both models was calculated. The more similar both approaches are, the higher the percentage of respondents that are identically classified will be.

Comparison #3. Similarity between the MIXL estimates

The third model comparison directly compared the MIXL estimates. In the MIXL model with a garbage class, the reported MIXL estimates are automatically corrected for the influence of respondents with low-quality response patterns. For the standard MIXL, however, four different sets of model estimates needed to be compared, i.e. estimates for the entire sample without any respondent selection as well as estimates for the subsets of good-quality respondents based on the three RLH cut-off values. As previously mentioned, when respondents with a low-quality response pattern are excluded from the sample, the MIXL estimates after respondent selection should have larger absolute mean values combined with a reduction in the relative size of the reported standard deviations – and become more similar to the estimates of the garbage class MIXL model.

RESULTS

Table 2 presents the percentage of respondents classified with a low-quality response pattern by the RLH statistic and the garbage class MIXL model. As shown, both methods produce close to identical classifications, with a mean and maximum difference of 2 and 5 percentage points, respectively, across the included datasets and scenarios.

Table 3 presents the percentage of respondents that are identically classified by both methods. Across all datasets and scenarios, approximately 95% of all respondents are identically classified by the RLH and garbage class MIXL models. As with the comparison at the aggregate level, the individual-level classification is slightly less congruent for scenario 1 (94%) than for scenarios 2 and 3 (96%) but the difference is small.

Table 4 provides a comparison of the MIXL model results for the antibiotics dataset, which is the smallest of the four datasets and easier to interpret than the EQ-5D table, which is included in the Appendix. As shown in Table 4, the inclusion of a garbage class has a major impact on the MIXL estimates. Most importantly on the choice consistency, with an approximately 2.5x increase in the size of the mean preference parameters. The size of the standard deviations of the normal distribution relative to the mean estimates also decreases. In the standard MIXL model, the SD estimates range from slightly smaller (0.9 times) to slightly larger (1.1 times) than the mean estimates, whereas in the garbage class MIXL model all SD estimates are somewhat (0.8 times) to substantially (0.5 times) smaller than the mean estimates. Finally, the relative attribute importances are also affected, with attributes 2 and 3 becoming somewhat (i.e. 13% and 18%) more important relative to attribute 1, respectively.

When comparing the garbage class MIXL estimates to the MIXL model estimates after respondents with a low RLH are excluded from the analyses, a similar effect can be observed. The more low-quality respondents are excluded, the stronger the MIXL estimates resemble the garbage class MIXL model estimates. Moreover, Tables 1-3 in Appendix A of the Online Supplemental provide the same sets of MIXL estimates for the other three datasets. Based on the smaller number of low-quality respondents in these datasets (cf. Table 2 and the garbage class share estimates), the difference in choice consistency, impact on the standard deviations of the normal distributions, and particularly shifts in relative attribute importance are smaller than in the antibiotics dataset. However, the same effects can be observed. In addition, the standard MIXL models after RLH selection produce close to identical estimates as the garbage class MIXL model.

Table 2. Percentage of respondents classified with a low-quality response pattern, by method, dataset, and cut-off values *

	scenario		
	#1	#2	#3
MIXL with RLH selection	**prob(RLH_i<1/J)>0.01**	**prob(RLH_i<1/J)>0.05**	**prob(RLH_i<1/J)>0.10**
Antibiotics	0.38	0.29	0.25
Vaccines	0.09	0.06	0.05
Meals	0.20	0.08	0.04
EQ-5D	0.27	0.14	0.10
MIXL with garbage class	**prob(φ_i<0.5)>0.20**	**prob(φ_i<0.5)>0.75**	**prob(φ_i<0.5)>0.95**
Antibiotics	0.37	0.31	0.27
Vaccines	0.12	0.09	0.08
Meals	0.26	0.08	0.03
EQ-5D	0.24	0.13	0.10
	absolute difference	**absolute difference**	**absolute difference**
Antibiotics	0.00	0.02	0.02
Vaccines	0.03	0.03	0.04
Meals	0.05	0.00	0.01
EQ-5D	0.03	0.01	0.01

** Note: scenario 1 = prob(RLH_i<1/J)>0.01 & prob(φ_i<0.5)>0.20,*
scenario 2 = prob(RLH_i<1/J)>0.05 & prob(φ_i<0.5)>0.75,
scenario 3 = prob(RLH_i<1/J)>0.10 & prob(φ_i<0.5)>0.95, with
RLH = root likelihood, J = number of choice options per choice task, and
φ_i = probability of MIXL (as opposed to garbage class) membership.

Table 3. Percentage of respondents identically classified by both methods, by dataset and cut-off value scenario *

	scenario *		
	1	**2**	**3**
Antibiotics	0.97	0.97	0.96
Vaccines	0.97	0.97	0.96
Meals	0.90	0.96	0.98
EQ-5D	0.93	0.94	0.94
Average	**0.94**	**0.96**	**0.96**

** Note: scenario 1 = prob(RLH$_i$<1/J)>0.01 & prob(φ_i<0.5)>0.20,*
scenario 2 = prob(RLH$_i$<1/J)>0.05 & prob(φ_i<0.5)>0.75,
scenario 3 = prob(RLH$_i$<1/J)>0.10 & prob(φ_i<0.5)>0.95, with
RLH = root likelihood, J = number of choice options per choice task, and
φ_i = probability of MIXL (as opposed to garbage class) membership.

Table 4. Antibiotics – Garbage class MIXL and MIXL estimates

	garbage MIXL		Standard MIXL		standard MIXL		standard MIXL		standard MIXL	
	N=750		N=750		N=466 *		N=533 **		N=563 ***	
	coef.	std.err.	coef.	std.err.	coef.	std.err.	coef.	std.err.	coef.	std.err.
speed (mean)	2.96	0.20	1.21	0.08	3.03	0.21	2.46	0.14	2.22	0.14
convenience (mean)	4.84	0.34	1.82	0.11	5.49	0.38	3.94	0.23	3.47	0.21
confidence (mean)	10.1	0.70	3.59	0.18	11.9	0.78	7.98	0.42	6.92	0.37
speed (SD)	1.58	0.18	1.10	0.09	1.59	0.18	1.29	0.13	1.21	0.16
convenience (SD)	3.81	0.32	2.01	0.11	4.55	0.38	3.26	0.23	2.97	0.21
confidence (SD)	5.91	0.50	3.33	0.18	6.96	0.59	4.92	0.36	4.48	0.34
garbage class share	0.34	0.01	n/a		n/a		n/a		n/a	

,,*** MIXL estimates after RLH selection with pr(RLH$_i$<½)>0.01, 0.05, and 0.10, respectively*

DISCUSSION

The proposed garbage class MIXL model is an elegantly simple extension of standard MIXL models. Hence it is unexpected that there are so few (if any) previous attempts to combine MIXL models with a garbage class, particularly because the impact of low-quality respondents on the MIXL model estimates was found to be substantial. Interestingly, the absence of existing applications is unrelated to the model's empirical performance: the garbage class MIXL model exhibits close to identical performance as the far more commonly used RLH statistic. Accordingly, the garbage class MIXL model represents a reliable method to accommodate for flat-lining, heuristics, and particularly random response patterns in the data, and has superior performance relative to internal validity tests such as repeated and dominant choice tasks.

From a practical perspective, the garbage class MIXL model can be easily implemented in existing software packages. The model is also easily generalizable to (stand-alone) DCE with duration data, which is of particular importance for the upcoming EQ-DCE-VT protocol. The garbage class MIXL also has the advantage of producing preferences estimates that are unaffected by low-quality respondents without having to manually screen for respondents with low data quality based on arbitrary cut-off values. The resulting shift in choice consistency and changes in relative attribute importances can have a profound impact on willingness-to-pay (WTP), maximum acceptable risk (MAR), DCE uptake predictions, and, in case of the EQ, estimated tariffs. In this sense the garbage class MIXL model provides a sensible default specification. That is, the garbage class MIXL automatically reduces to the standard MIXL if there are no respondents with low quality response patterns identified, yet provides MIXL estimates that are automatically purged from the impact of respondents with low data quality if such respondents do exist in the data.

Another important advantage of the garbage class MIXL model is that the garbage class membership probability estimates can be directly interpreted as measurements of DCE data quality.

- At the individual level, the estimated garbage class-membership probabilities can be used to identify respondents with low data quality in a very similar fashion as RLH selection. Similar to the RLH approach appropriate statistical cut-off values need to be selected, which implies that some degree of ambiguity remains in the classification of respondents with low data quality. In this respect, it is important to mention that the

RLH and garbage class approach were found to provide close to identical results. Therefore established sensitivity and specificity results of Jonker et al. [11] are also relevant to the garbage class MIXL model. More specifically, the $\text{prob}(RLH_i<1/J)>0.05$ reference classification rule closely corresponds to $\text{prob}(\varphi_i<0.5)>0.75$, which implies that the latter can be recommended for the garbage class MIXL model. Of course, more stringent cut-off values can be selected depending on the research objective. More importantly, in contrast to RLH selection, it is not necessary to re-fit the model when the subset of respondents with low-quality response data has been identified; the garbage class MIXL estimates already are purged from the influence of low-quality respondents.

- At the sample level, the average garbage class-membership probability summarizes the garbage class share and thus provides an estimate of the number of low-quality respondents in the dataset. As mentioned in the introduction, reliable indicators of DCE data quality are scarce whereas our field faces increasing pressure from policy makers, regulators, and other stakeholders to not only follow good research practices but to also ensure adequate quality control and to provide DCE data quality assurances. From this perspective, being able to objectively quantify DCE data quality based on an approach that is consistent with the underlying theoretical framework of DCEs is an important feature of the proposed model. Moreover, unlike individual-level respondent classification, garbage class membership probabilities are readily available from the model's output – without having to select arbitrary cut-off values.

Finally, as mentioned in the methods section, it is straight-forward to extend the garbage class MIXL specification with class-membership predictor variables. This could, for example, accommodate a formal evaluation of the determinants of garbage class-membership based on respondent characteristics and DCE response times. Such analyses were beyond the scope of this paper but constitute an interesting avenue for future research, e.g. to see whether DCE response times are accurate predictors of garbage class membership.

In terms of model flexibility, unlike (scale-adjusted) latent class logit models, the garbage class MIXL model does not assume identical within-class preferences. Hence the proposed model relaxes a restrictive assumption that in latent class logit models often results in the selection of too many classes, leading to over parameterization and many, relatively small, classes.[12] In contrast, the standard garbage class MIXL model is already quite flexible despite only comprising two classes. Of course, in some situations the standard model can be too parsimonious to adequately reflect the distribution of the good-quality respondents' preferences. In such cases, a more flexible mixing distribution would be warranted, potentially one that can accommodate multiple MIXL classes [22-24] but in case of EQ-5D related research preferably with a more flexible mixing distribution that does not automatically induce multi-modal mixing distributions.[25]

Another interesting comparison can be made between the garbage class MIXL model and the attribute non-attendance (ANA) literature. As mentioned by one of the reviewers of this paper, in an ANA framework the garbage class represents the extreme case when all of the attributes are ignored, and there could of course also be intermediate cases where some attributes are ignored but not all. In line with previous contributions modelling non-attendance in a mixed logit framework (e.g.[26-29]), accommodating such response styles in an ANA framework can be seen as an extension of the garbage class MIXL model, albeit at the cost of its appealing simplicity. This is certainly true. In addition, it should probably be mentioned that intermediate cases of non-attendance behavior can also be captured within a random heterogeneity specification like the garbage class MIXL model, which means that an ANA extension of the model is only recommendable when behavioral ANA estimates are of direct interest. In applied DCE research, it will often be preferable to rely on a garbage class MIXL model without ANA extensions, particularly when WTP or MAR or EQ-5D tariff estimates need to be reported.[28]

Even though statistical methods such as the garbage class MIXL model and RLH test statistics can relatively reliably detect low-quality respondents, they are unable to differentiate between those who a) are willing to give honest, thoughtful responses but truly do not care much about the included attributes, and b) those who are unmotivated, inattentive, and essentially provide dishonest answers to receive financial incentives with the least amount of effort. While the latter group of respondents should definitely be removed from the analyses, and particularly in online surveys are likely to represent the majority of garbage class respondents, it should be noted that the exclusion of the former group is undesirable and can also bias the estimates and uptake predictions.

Other disadvantages that are shared between the garbage class MIXL model and the RLH approach are that they both depend on the quality of the individual-level estimates. Hence their performance relies first of all on the correctness of the model specification but also on the efficiency of the DCE design and on the number of choice tasks per respondent. As such, both approaches benefit strongly from the use of efficient DCE designs that are optimized with informative, non-zero priors, particularly in the case of DCE with duration designs.[30] Vise versa, neither of the two approaches seems particularly recommendable if the number of choice tasks per respondent in the DCE is considerably smaller than the number of parameters in the utility function to be estimated.

As a final note, even though the garbage class MIXL model provides a convenient approach to be able to detect and remove the influence of respondents with low-quality response patterns from the statistical analyses of DCEs, also for DCE with duration datasets, it is neither intended nor recommended to be used as a substitute for a carefully designed survey instrument. After all, response quality and behavioral efficiency are not exogenously determined; they endogenously depend on respondents' engagement/motivation and on the level of task complexity, which in turn is affected by the DCE design dimensions and various design aspects, such as the type of experimental DCE design, the inclusion of attribute level overlap, the visual presentation of the choice tasks, and the inclusion of well-designed DCE warm-up tasks, see e.g. [31-34] The better the quality of the survey instrument and the more engaged and motivated the survey respondents are, the less important it is to fit a garbage class MIXL model. All of these aspects are already taken into consideration in the pilot EQ stand-alone DCE protocol. Accordingly, the optimal outcome is a garbage class MIXL model that assigns very few respondents to the garbage class and consequently produces almost identical results as the standard MIXL model, which would imply that the proposed model is merely used to ensure correct results and allows researchers to report that very few respondents were assigned to the garbage class.

REFERENCES

1. Pullenayegum E, Stolk E, Jakubczyck M, et al. Using Discrete Choice Experiments as a stand-alone approach to valuation: a draft protocol. In: Prague PpaEPMi, ed., 2020.
2. Himmler S, Jonker MF, van Krugten F, Hackert M, van Exel J, Brouwer W. Estimating an anchored utility tariff for the well-being of older people measure (WOOP) for the Netherlands. Social Science & Medicine. 2022 May 1;301:114901.
3. Prevolnik Rupel V, Ogorevc M. EQ-5D-Y value set for Slovenia. Pharmacoeconomics. 2021 Apr;39(4):463-71.
4. Shiroiwa T, Ikeda S, Noto S, Fukuda T, Stolk E. Valuation survey of EQ-5D-Y based on the international common protocol: development of a value set in Japan. Medical Decision Making. 2021 Jul;41(5):597-606.
5. Kreimeier S, Mott D, Ludwig K, Greiner W. EQ-5D-Y Value Set for Germany. PharmacoEconomics. 2022 May 23:1-3.
6. Ramos-Goñi JM, Oppe M, Estévez-Carrillo A, Rivero-Arias O, Wolfgang G, Simone K, Kristina L, Valentina R. Accounting for unobservable preference heterogeneity and evaluating alternative anchoring approaches to estimate country-specific EQ-5D-Y value sets: a case study using Spanish preference data. Value in Health. 2022 May 1;25(5):835-43.
7. Johnson FR, Yang JC, Reed SD. The internal validity of discrete choice experiment data: a testing tool for quantitative assessments. Value Health. 2019;22(2):157–160.
8. Janssen EM, Marshall DA, Hauber AB, Bridges JFP. Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability? Expert Rev Pharmacoecon Outcomes Res. 2017;17(6):531–542.
9. Tervonen T, Schmidt-Ott T, Marsh K, Bridges JFP, Quaife M, Janssen E. Assessing rationality in discrete choice experiments in health: an investigation into the use of dominance tests. Value Health. 2018;21(10):1192–1197.
10. Lancsar E, Louviere J. Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences? Health Econ. 2006;15(8):797–811.
11. Jonker MF, Roudijk B, Maas M. The Sensitivity and Specificity of Repeated and Dominant Choice Tasks in Discrete Choice Experiments. Value in Health. 2022
12. Soekhai V, de Bekker-Grob EW, Ellis AR, Vass CM. Discrete choice experiments in health economics: past, present and future. Pharmacoeconomics. 2019;37(2):201–226.
13. Magidson J, Vermunt JK. Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference. Sawtooth software conference 2007 Oct (Vol. 139).
14. Chrzan K, Halversen C. Diagnostics for random respondents in choice experiments. Sawtooth Software. https://sawtoothsoftware.com/resources/technicalpapers/diagnostics-for-random-respondents-in-choice-experiments. Accessed March 2022.

15. Hess S, Palma D. Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of Choice Modelling*, 32. doi: 10.1016/j.jocm.2019.100170.
16. Bierlaire M. BIOGEME: A free package for the estimation of discrete choice models. InSwiss transport research conference 2003 (No. CONF).
17. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. Statistics in medicine. 2009 Nov 10;28(25):3049-67.
18. Jonker MF, Roudijk B, Maas M. The Sensitivity and Specificity of Repeated and Dominant Choice Tasks in Discrete Choice Experiments. (2021) EuroQol Plenary
19. Lenk PJ, De Sarbo WS. Bayesian inference for finite mixtures of generalized linear models with random effects. Psychometrika. 2000 Mar;65(1):93-119.
20. De Blasi P, James LF, Lau JW. Bayesian nonparametric estimation and consistency of mixed multinomial logit choice models. Bernoulli. 2010 Aug;16(3):679-704.
21. Train KE. EM algorithms for nonparametric estimation of mixing distributions. Journal of Choice Modelling. 2008 Jan 1;1(1):40-69.
22. Greene WH, Hensher DA. Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. Applied Economics. 2013 May 1;45(14):1897-902.
23. Keane M, Wasi N. Comparing alternative models of heterogeneity in consumer choice behavior. Journal of Applied Econometrics. 2013 Sep;28(6):1018-45.
24. Zhou M, Bridges JF. Explore preference heterogeneity for treatment among people with type 2 diabetes: a comparison of random-parameters and latent-class estimation techniques. Journal of choice modelling. 2019 Mar 1;30:38-49.
25. Train K. Mixed logit with a flexible mixing distribution. Journal of choice modelling. 2016 Jun 1;19:40-53.
26. Scarpa R, Gilbride TJ, Campbell D, Hensher DA. Modelling attribute non-attendance in choice experiments for rural landscape valuation. European review of agricultural economics. 2009 Jun 1;36(2):151-74.
27. Hensher DA, Collins AT, Greene WH. Accounting for attribute non-attendance and common-metric aggregation in a probabilistic decision process mixed multinomial logit model: a warning on potential confounding. Transportation. 2013 Sep;40(5):1003-20.
28. Hess S, Stathopoulos A, Campbell D, O’Neill V, Caussade S. It’s not that I don’t care, I just don’t care very much: confounding between attribute non-attendance and taste heterogeneity. Transportation. 2013 May;40(3):583-607.
29. Jonker MF, Donkers B, de Bekker-Grob EW, Stolk EA. Effect of level overlap and color coding on attribute non-attendance in discrete choice experiments. Value in Health. 2018 Jul 1;21(7):767-71.
30. Jonker MF, Bliemer MC. On the optimization of Bayesian D-efficient discrete choice experiment designs for the estimation of QALY tariffs that are corrected for nonlinear time preferences. Value in Health. 2019 Oct 1;22(10):1162-9.
31. DeShazo JR, Fermo G. Designing choice sets for stated preference methods: the effects of complexity on choice consistency. Journal of Environmental Economics and management. 2002 Jul 1;44(1):123-43.

32. Caussade S, de Dios Ortúzar J, Rizzi LI, Hensher DA. Assessing the influence of design dimensions on stated choice experiment estimates. Transportation research part B: Methodological. 2005 Aug 1;39(7):621-40.

33. Bech M, Kjaer T, Lauridsen J. Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. Health economics. 2011 Mar;20(3):273-86.

34. Jonker MF, Donkers B, de Bekker-Grob E, Stolk EA. Attribute level overlap and color coding can reduce task complexity, improve choice consistency, and decrease the dropout rate in discrete choice experiments. Health economics. 2019;28(3):350-63

SUPPLEMENTAL

Appendix A - MIXL model estimates for the vaccines, meals, and EQ-5D studies

Table A1. Vaccines - Garbage class MIXL and MIXL estimates

	garbage MIXL		standard MIXL		standard MIXL		standard MIXL		standard MIXL	
	N=500		N=500		N=455 *		N=470 **		N=477 ***	
	coef.	std.err.	coef.	std.err.	coef.	std.err.	coef.	std.err.	coef.	std.err.
effectiveness 40% (mean)	3.23	0.22	2.21	0.17	2.90	0.23	2.66	0.21	2.51	0.18
effectiveness 60% (mean)	6.18	0.33	4.43	0.25	5.69	0.33	5.26	0.30	5.00	0.27
effectiveness 80% (mean)	8.41	0.43	6.17	0.33	7.87	0.42	7.28	0.38	6.93	0.35
severe effects 10/1m (mean)	-0.40	0.12	-0.37	0.10	-0.38	0.12	-0.40	0.11	-0.38	0.12
severe effects 100/1m (mean)	-1.82	0.19	-1.50	0.16	-1.82	0.19	-1.72	0.17	-1.65	0.17
mild effects 3/10 (mean)	-0.18	0.11	-0.12	0.09	-0.16	0.11	-0.18	0.10	-0.16	0.09
mild effects 5/10 (mean)	-0.69	0.12	-0.53	0.10	-0.65	0.11	-0.65	0.11	-0.61	0.11
duration 6m (mean)	0.63	0.12	0.48	0.10	0.60	0.12	0.55	0.10	0.55	0.10
duration 12m (mean)	1.48	0.14	1.12	0.11	1.43	0.13	1.33	0.12	1.28	0.11
protection after 4w (mean)	-0.18	0.09	-0.12	0.08	-0.19	0.09	-0.15	0.09	-0.13	0.08
no vaccine (mean)	-2.89	0.96	-1.10	0.46	-2.38	0.78	-1.68	0.63	-1.47	0.57
effectiveness 40% (SD)	2.24	0.21	1.93	0.16	2.12	0.21	2.10	0.19	2.07	0.17
effectiveness 60% (SD)	4.41	0.30	3.84	0.22	4.23	0.29	4.12	0.26	4.04	0.25
effectiveness 80% (SD)	6.25	0.38	5.47	0.29	5.99	0.37	5.82	0.32	5.74	0.32
severe effects 10/1m (SD)	1.05	0.13	0.89	0.12	1.08	0.14	1.03	0.13	0.96	0.13
severe effects 100/1m (SD)	2.54	0.20	2.15	0.17	2.54	0.20	2.44	0.19	2.35	0.18
mild effects 3/10 (SD)	0.62	0.12	0.50	0.09	0.55	0.11	0.57	0.10	0.53	0.10
mild effects 5/10 (SD)	1.07	0.15	0.82	0.13	0.89	0.13	0.94	0.13	0.91	0.13
duration 6m (SD)	0.71	0.13	0.70	0.10	0.78	0.12	0.75	0.12	0.72	0.10
duration 12m (SD)	1.34	0.15	1.28	0.12	1.38	0.14	1.35	0.14	1.30	0.12
protection after 4w (SD)	0.51	0.11	0.48	0.09	0.53	0.10	0.54	0.10	0.53	0.10
no vaccine (SD)	15.0	1.16	8.03	0.56	12.7	1.00	10.5	0.81	9.61	0.68

garbage class share	0.11	0.01	n/a	n/a	n/a	n/a

*,**,*** MIXL estimates after RLH selection with $pr(RLH_i<⅓)>0.01$, 0.05, and 0.10, respectively

Table A2. Meals - Garbage class MIXL and MIXL estimates

	garbage MIXL		standard MIXL		standard MIXL		standard MIXL		standard MIXL	
	N=500		N=500		N=400 *		N=459 **		N=478 ***	
	coef.	std.err.	coef.	std.err.	coef.	std.err.	coef.	std.err.	coef.	std.err.
tastes good (mean)	1.72	0.14	1.36	0.10	2.08	0.17	1.49	0.12	1.51	0.12
tastes very good (mean)	3.55	0.20	2.66	0.15	4.06	0.24	3.11	0.18	2.94	0.17
health neutral (mean)	3.27	0.20	2.43	0.15	3.70	0.24	2.96	0.19	2.62	0.17
healthy (mean)	5.01	0.27	3.74	0.19	5.65	0.32	4.53	0.24	4.09	0.22
15 min prep. time (mean)	-0.34	0.10	-0.18	0.07	-0.40	0.10	-0.28	0.09	-0.30	0.09
30 min prep. time (mean)	-1.07	0.12	-0.67	0.08	-1.18	0.12	-0.84	0.10	-0.85	0.10
45 min prep. time (mean)	-2.07	0.15	-1.50	0.11	-2.39	0.18	-1.82	0.14	-1.76	0.13
10 min travel time (mean)	-0.30	0.08	-0.18	0.06	-0.31	0.09	-0.22	0.08	-0.17	0.08
20 min travel time (mean)	-0.95	0.09	-0.64	0.07	-1.06	0.10	-0.72	0.09	-0.73	0.08
30 min travel time (mean)	-1.51	0.10	-1.03	0.07	-1.54	0.11	-1.18	0.09	-1.12	0.09
$4 price (mean)	-0.43	0.10	-0.24	0.07	-0.36	0.10	-0.32	0.09	-0.33	0.08
$6 price (mean)	-1.08	0.13	-0.68	0.10	-1.15	0.15	-0.90	0.12	-0.81	0.11
$8 price (mean)	-2.28	0.18	-1.66	0.13	-2.56	0.20	-1.99	0.17	-1.87	0.15
no meal (mean)	-3.34	0.58	-2.73	0.35	-2.40	0.46	-3.01	0.45	-3.14	0.40
tastes good (SD)	1.36	0.15	1.24	0.10	1.51	0.17	1.29	0.13	1.26	0.12
tastes very good (SD)	2.67	0.21	2.50	0.14	3.03	0.23	2.64	0.17	2.66	0.16
health neutral (SD)	2.40	0.21	2.34	0.14	2.82	0.25	2.62	0.18	2.42	0.17
healthy (SD)	3.68	0.27	3.40	0.18	4.08	0.32	3.72	0.23	3.56	0.22
15 min prep. time (SD)	0.70	0.14	0.69	0.09	0.84	0.16	0.76	0.11	0.68	0.11
30 min prep. time (SD)	1.22	0.14	1.15	0.09	1.41	0.16	1.22	0.12	1.15	0.11
45 min prep. time (SD)	1.83	0.17	1.75	0.11	2.28	0.22	1.93	0.15	1.78	0.15
10 min travel time (SD)	0.53	0.10	0.46	0.07	0.56	0.11	0.60	0.10	0.47	0.08
20 min travel time (SD)	0.70	0.10	0.67	0.08	0.74	0.13	0.77	0.10	0.71	0.09

30 min travel time (SD)	0.91	0.12	0.91	0.09	0.97	0.16	0.95	0.11	0.81	0.09
$4 price (SD)	0.85	0.13	0.81	0.08	0.85	0.12	0.91	0.11	0.65	0.09
$6 price (SD)	1.85	0.15	1.55	0.10	1.98	0.18	1.71	0.14	1.49	0.12
$8 price (SD)	2.77	0.20	2.44	0.13	3.08	0.21	2.69	0.17	2.46	0.15
no meal (SD)	5.35	0.60	4.33	0.33	5.24	0.51	5.20	0.45	4.70	0.40
garbage class share	0.16	0.01	n/a		n/a		n/a		n/a	

*,**,*** MIXL estimates after RLH selection with $pr(RLH_i<⅕)>0.01$, 0.05, and 0.10, respectively

Table A3. EQ-5D - Garbage class MIXL and MIXL estimates

	garbage MIXL		standard MIXL		standard MIXL		standard MIXL		standard MIXL	
	N=500		N=500		N=365 *		N=428 **		N=451 ***	
	coef.	std.err.	coef.	std.err.	coef.	std.err.	coef.	std.err.	coef.	std.err.
mobility 2 (mean)	-0.82	0.11	-0.71	0.08	-0.80	0.12	-0.80	0.10	-0.77	0.10
mobility 3 (mean)	-1.61	0.13	-1.30	0.10	-1.59	0.13	-1.49	0.11	-1.48	0.11
mobility 4 (mean)	-2.92	0.16	-2.34	0.13	-2.93	0.16	-2.73	0.15	-2.63	0.15
mobility 5 (mean)	-4.70	0.22	-3.70	0.17	-4.86	0.22	-4.42	0.20	-4.22	0.20
usual activities 2 (mean)	-0.65	0.12	-0.55	0.08	-0.7	0.12	-0.65	0.11	-0.59	0.10
usual activities 3 (mean)	-1.31	0.12	-1.06	0.09	-1.37	0.13	-1.27	0.11	-1.19	0.10
usual activities 4 (mean)	-2.80	0.16	-2.20	0.12	-2.99	0.17	-2.66	0.14	-2.50	0.13
usual activities 5 (mean)	-3.99	0.21	-3.08	0.16	-4.27	0.23	-3.75	0.19	-3.53	0.19
self-care 2 (mean)	-0.46	0.11	-0.39	0.09	-0.35	0.12	-0.49	0.10	-0.44	0.10
self-care 3 (mean)	-1.18	0.13	-0.96	0.10	-1.14	0.14	-1.18	0.11	-1.08	0.10
self-care 4 (mean)	-2.33	0.15	-1.86	0.12	-2.37	0.17	-2.25	0.13	-2.11	0.14
self-care 5 (mean)	-3.36	0.18	-2.70	0.14	-3.48	0.21	-3.24	0.17	-3.07	0.17
pain & discomfort 2 (mean)	-0.91	0.11	-0.74	0.09	-0.9	0.13	-0.85	0.11	-0.80	0.10
pain & discomfort 3 (mean)	-2.20	0.14	-1.70	0.11	-2.27	0.14	-2.03	0.13	-1.93	0.13
pain & discomfort 4 (mean)	-4.21	0.21	-3.31	0.16	-4.46	0.22	-3.96	0.20	-3.71	0.19
pain & discomfort 5 (mean)	-5.95	0.29	-4.61	0.22	-6.29	0.29	-5.55	0.27	-5.26	0.26
anxiety & depression 2 (mean)	-0.78	0.11	-0.67	0.09	-0.75	0.13	-0.72	0.10	-0.71	0.11
anxiety & depression 3 (mean)	-1.97	0.14	-1.60	0.11	-1.99	0.14	-1.84	0.12	-1.79	0.13
anxiety & depression 4 (mean)	-3.81	0.20	-3.04	0.15	-4.00	0.22	-3.54	0.18	-3.44	0.18
anxiety & depression 5 (mean)	-5.07	0.27	-3.99	0.19	-5.27	0.28	-4.69	0.24	-4.50	0.24
mobility 2 (SD)	0.51	0.10	0.50	0.09	0.53	0.11	0.52	0.10	0.48	0.09
mobility 3 (SD)	0.64	0.11	0.84	0.11	0.64	0.14	0.70	0.12	0.65	0.10
mobility 4 (SD)	1.11	0.15	1.43	0.14	1.14	0.17	1.21	0.16	1.20	0.16

mobility 5 (SD)	1.68	0.17	2.35	0.17	1.85	0.24	1.91	0.20	2.04	0.21
usual activities 2 (SD)	0.54	0.12	0.45	0.09	0.54	0.12	0.53	0.12	0.48	0.10
usual activities 3 (SD)	0.66	0.12	0.64	0.11	0.73	0.12	0.63	0.11	0.59	0.10
usual activities 4 (SD)	1.06	0.17	1.24	0.14	1.13	0.17	1.13	0.18	1.13	0.16
usual activities 5 (SD)	2.01	0.20	2.22	0.17	2.28	0.23	2.31	0.21	2.26	0.19
self-care 2 (SD)	0.55	0.11	0.51	0.09	0.51	0.10	0.55	0.11	0.49	0.09
self-care 3 (SD)	0.55	0.11	0.57	0.12	0.49	0.10	0.50	0.11	0.51	0.10
self-care 4 (SD)	0.72	0.13	1.06	0.15	0.88	0.17	0.91	0.15	1.06	0.15
self-care 5 (SD)	1.18	0.17	1.68	0.18	1.55	0.20	1.53	0.19	1.73	0.17
pain & discomfort 2 (SD)	0.61	0.12	0.62	0.09	0.65	0.11	0.56	0.10	0.55	0.11
pain & discomfort 3 (SD)	0.98	0.14	1.17	0.12	1.06	0.15	0.91	0.14	0.97	0.14
pain & discomfort 4 (SD)	1.94	0.19	2.06	0.16	2.24	0.21	1.83	0.22	2.00	0.20
pain & discomfort 5 (SD)	2.81	0.23	3.04	0.19	3.27	0.29	2.75	0.27	2.96	0.25
anxiety & depression 2 (SD)	0.70	0.12	0.56	0.10	0.71	0.12	0.62	0.11	0.55	0.10
anxiety & depression 3 (SD)	1.01	0.13	1.00	0.11	1.05	0.16	0.95	0.13	0.83	0.13
anxiety & depression 4 (SD)	1.92	0.19	1.88	0.14	2.14	0.22	1.92	0.17	1.78	0.17
anxiety & depression 5 (SD)	2.63	0.24	2.65	0.19	3.02	0.28	2.63	0.23	2.55	0.23
garbage class share	0.19	0.01	n/a		n/a		n/a		n/a	

*,**,*** MIXL estimates after RLH selection with pr(RLH_i<½)>0.01, 0.05, and 0.10, respectively