Methodology

The Garbage Class Mixed Logit Model: Accounting for Low-Quality Response Patterns in Discrete Choice Experiments

Marcel F. Jonker, PhD

ABSTRACT

Objectives: To introduce the garbage class mixed logit (MIXL) model as a convenient alternative to manually screening and accounting for respondents with low data quality in discrete choice experiments.

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 MIXL models as well.

Results: Based on a reanalysis of 4 discrete choice experiments that were originally analyzed using a standard MIXL model, it is shown that garbage class MIXL models can achieve the same effect as manually screening for (and excluding) respondents with low data quality based on the more commonly used root likelihood test, but with less effort and ambiguity.

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 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 an attractive alternative.

Keywords: data quality, discrete choice experiment, garbage class, mixed logit model.

Introduction

Internal validity tests are checks on the logic, consistency, and trade-off assumptions in discrete choice data and are often used to assess respondents’ response quality in discrete choice experiments (DCEs). In about 30% of all recently published health-related DCEs, 1 or more internal validity tests are included. Interestingly, the use of internal validity tests is substantially less prevalent in other scientific fields, such as marketing (7%), environmental sciences (9%), and transport economics (10%). This difference is indicative of the enormous demand for the verification and quality control of response data in health preference elicitations, for example, as part of ongoing efforts to make medical, regulatory, and health technology assessments more patient-centered. It also reflects a certain degree of inertia (ie, practitioners tend to copy methods that are used by others), the availability of dedicated software, a scarcity of suitable alternatives, and certainly also a degree of ignorance concerning the theoretical and practical problems associated with internal validity tests.

From a theoretical perspective, internal validity tests cannot take response error into account and are consequently inconsistent with the theoretical foundation of DCEs. Hence, violations of internal validity tests are notoriously difficult to interpret, particularly when the predicted utility difference between the included choice options is small. 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. As mentioned by Jonker et al. 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. More importantly, also from an empirical perspective, the performance of internal validity tests appears to be inadequate. For example, the predictive accuracy of the 2 most commonly used internal validity tests, repeated and dominant choice tasks, is not much better than a random coin flip. In contrast, likelihood-based statistical validity tests, such as the root likelihood (RLH) statistic, provide a superior alternative for the assessment of respondents’ response quality in DCEs.

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 1 or more statistical cutoff
values to be able to classify respondents as either having provided good or bad quality responses. Subsequently, assuming that 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 1 or more statistical cutoff 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 article, 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 2 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 logit models as introduced by Magidson and Vermunt, particularly those in which 1 of the scale classes has a scale constrained to 0. Nevertheless, 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 1 minus that probability to the garbage class; this is based on the match between the respondents’ response patterns and the 2 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, that is, without having to manually conduct split sample analyses based on internal or statistical validity tests.

In the remainder of this article, 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 4 different datasets that were previously analyzed using standard MIXL models, the garbage class MIXL model is presented as a simpler alternative to manually screening for respondents with low data quality while automatically providing MIXL estimates that are unaffected by the presence of respondents with low-quality response patterns.

### 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 as

\[ X_{ijk} (r) \text{ for } i = 1, \ldots, N; \ t = 1, \ldots, T; \ j = 1, \ldots, J; \ k = 1, \ldots, K \]  

and all observed choices as

\[ Y_{ij} \in \{0, 1\}, \]  

with the dependent variable (Y) being equal to 1 for the alternative that was chosen and 0 for all other alternatives in each choice task, that is,

\[ (Y_{ij} = 1) \Rightarrow (Y_{im} = 0, \forall m \neq j). \]  

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

\[ U_{ij} = V_{ij} + \epsilon_{ij} \]  

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

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

\[ V_{ij} = \sum_{k=1}^{K} \beta_{ik} \ast X_{ijk} \]  

with \( \beta_{i} \) denoting a vector of K coefficients that can take any value across respondents and with \( \theta \) denoting the coefficients of the joint distribution.

2. **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

\[ P_{ij} = \frac{\exp (V_{ij})}{\sum_{k=1}^{K} \exp (V_{ik})} \]  

In the standard MIXL model, the mixing distribution of the respondents’ \( \beta \)-coefficients is assumed to be multivariate normal (MVN) with mean vector \( \mu \) and covariance matrix \( \Sigma \), that is

\[ f(\beta) \sim MVN(\mu, \Sigma). \]  

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

\[ L_t = (2\pi)^{-\frac{1}{2}} |\det (\Sigma)|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} \left( (\beta_i - \mu) \Sigma^{-1} (\beta_i - \mu) \right) \right] \]  

\[ + \sum_{t=1}^{T} \sum_{j=1}^{J} (Y_{ij} \log (P_{ij})) \int d\beta_i \]
and the individual-level RLH statistic is defined as the geometric mean of the respondents’ likelihood across the T choice tasks in the DCE

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

Interestingly, a null model with equal choice probabilities for all choice options (ie, based on entirely random choice patterns) results in a RLH of 1/J. As such, respondents with RLH ≤ 1/J can 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 ≤ 1/J and to compare this statistic with different cutoff values (eg, 0.01, 0.05, and 0.10). If the estimated probability is larger than the chosen cutoff 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 \((V_{ij})\) with a class membership parameter \((\phi_i)\)

$$U_{ij} = \phi_i \cdot V_{ij} + \epsilon_{ij}.$$  \(10\)

No other changes are necessary and the interpretation of the class membership parameter is also straight-forward: \(\phi_i\) represents the respondent-specific probability of being assigned to the standard MIXL utility specification; if \(\phi_i = 1\), respondents are entirely assigned to the standard MIXL utility specification \((U_{ij} = V_{ij} + \epsilon_{ij})\), and if \(\phi_i = 0\) there is no contribution from the structural part of the utility function and respondents make choices entirely based on the error term \((U_{ij} = \epsilon_{ij})\). When aggregated, the sample average \(\phi_i\) represents the class share of the MIXL model and the sample average of \(1 - \phi_i\) 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:

$$\phi_i = \frac{\exp(\gamma Z_i)}{1+\exp(\gamma Z_i)},$$  \(11\)

where \(Z_i\) and \(\gamma\) 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 \(\phi_i\) parameters in this article 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 parameter 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 toward 0 with a corresponding increase in the relative size of SDs 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 cutoff 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 (eg, using Apollo\(^{10}\) or Biogeme\(^{12}\) software), but they 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 was used in this article. Uninformative MVN priors (ie, with a mean of 0 and SD of 10) were assigned to \(\mu\), Bernoulli(0.5) priors to the \(\phi_i\) parameters and a Wishart prior with an identity scale matrix and K degrees of freedom to the inverse variance-covariance matrix (ie, \(\Sigma^{-1}\)). Accordingly, the MIXL specification allows for potentially correlated preference parameters. Standard Gibbs update steps were used to update \(\mu\) and \(\Sigma^{-1}\), slice update steps to update \(\phi_i\), and a Metropolis-within-Gibbs algorithm with antithetic sampling as described by Bédard et al\(^{11}\) (2014) to update the \(\beta_i\) parameters. All estimations were performed using the OpenBUGS software package\(^{12}\) 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 reanalysis of 4 health-related DCEs. In a previous publication, 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. 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 4 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 Amazon Mechanical Turk surveys. This ensured a mixture of good and bad quality respondents, which allows 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, and 75+ years). Appendix C in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.01.015 provides a detailed overview of the sample representativeness and survey dropout rates in each of the 4 datasets.

**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 4 datasets. For the standard MIXL model, based on Jonker et al,\(^{13}\) 3 increasingly more conservative cutoff values were specified based on the mean posterior probability of the respondents’ RLH < 1/J (ie, low-response quality) being larger than 0.01, 0.05, and 0.10, respectively. For the garbage class MIXL model, similar
cutoff values were specified based on the mean posterior probability of \( \phi_i < 0.50 \) (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 4 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 3 increasingly more conservative cutoff 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, 4 different sets of model estimates needed to be compared, that is, estimates for the entire sample without any respondent selection as well as estimates for the subsets of good-quality respondents based on the 3 RLH cutoff 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 SDs—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 1. Overview of the 4 datasets that were used in the MIXL model comparisons.

| DCE topic* | DCE design | Number of attributes/levels | Optout | Number of parameters | Number of respondents | Number of choice tasks per respondent | Number of choice options per choice task |
|------------|------------|-----------------------------|--------|----------------------|-----------------------|--------------------------------------|------------------------------------------|
| 1. Antibiotics7 | Full factorial7 | 2/2/2 | No | 3 | 750 | 13 | 2 |
| 2. Vaccines8 | Bayesian D-efficient8 | 4/3/3/2 | Yes | 11 | 500 | 16 | 3 |
| 3. Meals9 | Bayesian D-Efficient9 | 3/3/4/4/4 | Yes | 14 | 500 | 14 | 5 |
| 4. EQ-5D-5L10 | Bayesian D-Efficient10 | 5/5/5/5/5 | No | 20 | 500 | 21 | 2 |

DCE indicates discrete choice experiment; EQ-5D-5L, 5-level version of EQ-5D.

* Citations of the original publications.

†Each respondent completed all possible non-dominant and 2 dominant plus 2 repeated choice tasks.

‡Each respondent completed 1 subdesign of a heterogeneous Bayesian DCE design with 10 subdesigns.13

### Table 2. Percentage of respondents classified with a low-quality response pattern based on method, dataset, and cutoff values.

| Method/Dataset | Scenario | 1 | 2 | 3 |
|----------------|----------|---|---|---|
| MIXL with RLH selection | prob(\(RLH_i<1/J\)) > 0.01 | 0.38 | 0.29 | 0.25 |
| Antibiotics | | 0.09 | 0.06 | 0.05 |
| Vaccines | | 0.20 | 0.08 | 0.04 |
| Meals | | 0.27 | 0.14 | 0.10 |
| EQ-5D | | | | |
| MIXL with garbage class | prob(\(\phi_i<0.5\)) > 0.20 | 0.37 | 0.31 | 0.27 |
| Antibiotics | | 0.12 | 0.09 | 0.08 |
| Vaccines | | 0.26 | 0.08 | 0.03 |
| Meals | | 0.24 | 0.13 | 0.10 |
| EQ-5D | | | | |
| Difference | Absolute difference | 0.00 | 0.02 | 0.02 |
| Antibiotics | | 0.03 | 0.03 | 0.04 |
| Vaccines | | 0.05 | 0.00 | 0.01 |
| Meals | | 0.03 | 0.01 | 0.01 |
| EQ-5D | | | | |

J indicates number of choice options per choice task; MIXL, mixed logit; RLH, root likelihood; \(\phi_i\), probability of MIXL (as opposed to garbage class) membership.
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 4 datasets. As shown, 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 SDs 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.

When comparing the garbage class MIXL estimates with the MIXL 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 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.07.013 provide the same sets of MIXL estimates for the other 3 datasets. Based on the smaller number of low-quality respondents in these datasets (cf. Table 3 and the garbage class share estimates), the difference in choice consistency, impact on the SDs of the normal distributions, and shifts in relative attribute importance are smaller than in the antibiotics dataset. Nevertheless, the exact same effects can be observed and the standard MIXL models after RLH selection again produce close to identical estimates as the garbage class MIXL model.

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. It 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 cutoff values. The resulting shift in choice consistency and changes in relative attribute importance can have a particularly profound impact on willingness-to-pay, maximum acceptable risk, and DCE uptake predictions. Their estimates are often relevant from a policy perspective, which also implies that 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 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.

Table 4. Antibiotics: Garbage class MIXL and MIXL estimates.

|                  | Garbage MIXL   | Standard MIXL   |
|------------------|----------------|-----------------|
|                  | N = 750        | N = 750         |
| Coef. SE         | Coef. SE       | Coef. SE        | Coef. SE       |
| Speed (mean)     | 2.96 0.20      | 1.21 0.08       | 3.03 0.21      | 2.46 0.14       |
| Convenience (mean)| 4.84 0.34     | 1.82 0.11       | 5.49 0.38      | 3.94 0.23       |
| Confidence (mean)| 10.1 0.70      | 3.59 0.18       | 11.9 0.78      | 7.98 0.42       |
| Speed (SD)       | 1.58 0.18      | 1.10 0.09       | 1.59 0.18      | 1.29 0.13       |
| Convenience (SD) | 3.81 0.32      | 2.01 0.11       | 4.55 0.38      | 3.26 0.23       |
| Confidence (SD)  | 5.91 0.50      | 3.33 0.18       | 6.96 0.59      | 4.92 0.36       |
| Garbage class share| 0.34 0.01      | N/A             | 0.69 0.59      | 4.48 0.34       |

Coef. indicates coefficient; MIXL, mixed logit; N/A, not applicable; SE, standard error.

*1/J, MIXL estimates after RLH selection with prob(RLH < 1/J) > 0.01, 0.05, and 0.10, respectively.
1. 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 cutoff 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.14 are also relevant to the garbage class MIXL model. More specifically, the \( \text{prob}(\text{RLH}_i < 0.1) \) > 0.05 reference classification rule closely corresponds to \( \text{prob}(\phi_i < 0.5) > 0.75 \), which implies that the latter can be recommended for the garbage class MIXL model. Of course, more stringent cutoff 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.

2. 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 cutoff values.

Finally, as mentioned in the methods section, it is straightforward 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 article but constitute an interesting avenue for future research.

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.16 In contrast, the standard garbage class MIXL model is already quite flexible despite only comprising 2 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.17–18 These mixed-MIXL models do tend to require larger sample sizes and more choice tasks per respondent for the adequate identification of the number of classes and respondent-specific class membership parameters. In applied health preference research, which is frequently based on relatively small sample sizes, such models are therefore not by definition an improvement. But if they can be fitted, they have the advantage of including both the latent class logit specification (ie, when the within-class variances go to 0) and the standard MIXL specification (ie, when a single MIXL class is found to be optimal) as a special case.

Another interesting comparison can be made between the garbage class MIXL model and the attribute nonattendance (ANA) literature. As mentioned by one of the reviewers of this article, 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 modeling nonattendance in a MIXL framework,19–22 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 nonattendance behavior can also be captured within a random heterogeneity specification, such as 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 willingness-to-pay or maximum acceptable risk estimates need to be reported.23

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 (1) are willing to give honest, thoughtful responses but truly do not care much about the included attributes, and (2) those who are unmotivated, inattentive, and essentially provide dishonest answers to receive financial incentives with the least amount of effort. Although the latter group of respondents should definitely be removed from the analyses, and particularly in online surveys likely 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 not only 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, nonzero priors. Vice versa, neither of the 2 approaches seem 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, 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.23–26 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. 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.
Supplemental Materials

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2022.07.013.

Article and Author Information

Accepted for Publication: July 13, 2022
Published Online: October 3, 2022
doi: https://doi.org/10.1016/j.jval.2022.07.013

Author Affiliations: Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, The Netherlands (Jonker); Erasmus Choice Modelling Centre, Erasmus University Rotterdam, Rotterdam, The Netherlands (Jonker).

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

Author Contributions: Concept and design: Jonker
Acquisition of data: Jonker
Analysis and interpretation of data: Jonker
Drafting of the manuscript: Jonker
Statistical analysis: Jonker
Obtaining funding: Jonker

Conflict of Interest Disclosures: Dr Jonker reported being a member of the EuroQol Group. No other disclosures were reported.

Funding/Support: This work was supported by a grant (EQ Project 1413-Ra) from the EuroQol Research Foundation.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Jonker MF, Roudijk B, Maas M. The sensitivity and specificity of repeated and dominant choice tasks in discrete choice experiments. Value Health. 2022. https://doi.org/10.1016/j.jval.2022.01.015.
2. Janssen EM, Marshall DA, Hauber AR, 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.
3. Johnson FR, Yang J-C, Reed SD. The internal validity of discrete choice experiment data: a testing tool for quantitative assessments. Value Health. 2019;22(2):157–160.
4. Lancsar E, Louviere J. Deleting ‘irrational’ responses from discrete choice experiments: a case of investigating or imposing preferences? Health Econ. 2006;15(8):757–761.
5. Tervonen T, Schmidt-Ott T, Marsh K, Bridges JFP, Quafe 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.
6. 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.
7. Magidson J, Vermunt JK. Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference. Sawtooth Software Conference Proceedings Vol 139: Sequim, WA; 2007.
8. Chrzan K, Halversen C. Diagnostics for random respondents in choice experiments. Sawtooth Software. https://content.sawtoothsoftware.com/assets/45eb9c45-78d5-435b-b7d0-3a350b86adad1. Accessed August 29, 2022.
9. Hess S, Palma D. Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application. J Choice Modell. 2019;12:100170.
10. Bierlaire M. BIOGEME: a free package for the estimation of discrete choice models. In: Swiss Transport Research Conference; Monte Verita, Switzerland. 2003.
11. Bédard M, Douc R, Moulines E. Scaling analysis of delayed rejection MCMC methods. Methodol Comput Appl Probab. 2014;16(4):811–838.
12. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: evolution, critique and future directions. Stat Med. 2009;28(25):3049–3067.
13. Såndor Z, Wedel M. Heterogeneous conjoint choice designs. J Mark Res. 2005;42(2):210–218.
14. Lenk PJ, DeSarbo WS. Bayesian inference for discrete choice models with random effects. Psychometrika. 2000;65(1):93–119.
15. De Blasi P, James LF, Lau JW. Bayesian nonparametric estimation and consistency of mixed multinomial logit choice models. Bernoulli. 2010;16(3):679–704.
16. Greene WH, Hensher DA. Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. Appl Econ. 2013;45(14):1897–1902.
17. Keane M, Wasi N. Comparing alternative models of heterogeneity in consumer choice behavior. J Appl Econ. 2013;28(6):1018–1045.
18. Zhou M, Bridges JFP. Explore preference heterogeneity for treatment among people with type 2 diabetes: a comparison of random-parameters and latent-class estimation techniques. J Choice Modell. 2019;30:38–49.
19. Scarpa R, Gilbride TJ, Campbell D, Hensher DA. Modelling attribute non-attendance in choice experiments for rural landscape valuation. Eur Rev Agric Econ. 2009;36(2):151–174.
20. 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;40(3):1003–1020.
21. Hess S, Stathopoulou 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;40(3):583–607.
22. 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 Health. 2018;21(7):767–771.
23. DeShazo JR, Fermo G. Designing choice sets for stated preference methods: the effects of complexity on choice consistency. J Environ Econ Manag. 2002;44(1):123–143.
24. Caussade S, de Dios Ortúzar J, Rizzi Ll, Hensher DA. Assessing the influence of design dimensions on stated choice experiment estimates. Transp Res B Methodol. 2005;39(7):621–640.
25. Bech M, Kjaer T, Lauridsen J. Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. Health Econ. 2011;20(3):273–286.
26. 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 Econ. 2019;28(3):350–363.