Journal of Business Economics
https://doi.org/10.1007/s11573-025-01241-y

ORIGINAL RESEARCH PAPER

Identifying nested preference structures in choice models based on stated choice data

Nils Goeken[1] · Peter Kurz[2] · Winfried J. Steiner[1]

Received: 5 April 2024 / Accepted: 6 September 2025

Abstract

Nowadays, choice modelling based on stated choice data is widely used for assessing consumer preferences, with the hierarchical Bayes multinomial logit (HB-MNL) model being typically embedded as choice model for preference estimation. The HB-MNL model accounts for individual preference heterogeneity and is therefore able to weaken the Independence of Irrelevant Alternatives property (IIA) the standard MNL model suffers from. The IIA property states that the ratio of choice probabilities of two alternatives remains constant independent of the availability of further alternatives and hence implies an often unrealistic proportional substitution pattern among alternatives. The IIA property can also be addressed by applying nested multinomial logit (NMNL) models, which explicitly allow for a representation of different degrees of similarity between subsets of alternatives (nests). In this paper, we consider a HB-NMNL model which combines both the idea of nested market structures and heterogeneous consumer preferences with the expectation to handle the IIA property even better. We propose an extensive simulation study in which we compare the performance of the established HB-MNL model to the HB-NMNL model under varying experimental conditions for model fit, prediction accuracy, and parameter recovery. Our findings from this Monte Carlo study are twofold: first, both types of models perform rather close with regard to predictive validity (and model fit). This result is quite surprising and provides strong support for the use of the less complex HB-MNL model. Second, as could be expected, the HB-NMNL model shows advantages in terms of parameter recovery which is an important criterion for product design decision-making. We further present results from applying and comparing both models in an empirical study for summer tires. Model estimation is carried out using the publicly available *R* software package *RSGHB*.

Keywords Stated choice data · Consumer heterogeneity · Nested multinomial logit · Hierarchical Bayes · Simulation

JEL Classification M31 · C11 · C15

Extended author information available on the last page of the article

Published online: 15 December 2025

1 Introduction

Choice modelling is nowadays the most widely used method for assessing consumer preferences based on stated choice data. Choice modelling for analyzing stated choice data goes back to Louviere and Woodworth (1983) who linked conjoint analysis for the first time to discrete choice modeling.[1] Here, in contrast to traditional conjoint analysis, preferences for stimuli are collected through choice decisions rather than by ranking or rating tasks. Since the seminal paper of Louviere and Woodworth (1983), the marketing literature has seen a tremendous number of choice modelling applications and related methodical advances for collecting and analyzing stated choice data. The primary reason for the increasing dominance of discrete choice experiments over the last 35 years is that they closely mimic real choice behavior of consumers by asking respondents repeatedly to choose their preferred alternative from a set of several offered alternatives (choice sets), respectively. Within the framework of choice modelling the multinomial logit (MNL) model is the most frequently used discrete choice model due to the existence of closed form solutions for conditional choice probabilities (e.g., Allenby et al. 1998; Gensler et al. 2012; Hein et al. 2019a).

There are two main areas of estimating choice models based on stated choice data beyond the pure measurement and analysis of consumer preferences. Both are based on the estimated preferences in the first step (i.e., the model estimation stage) and are connected to the decision-making of firms. The one is preference or market simulation in which expected utilities, expected individual consumer choices, or aggregate preference shares as indicators of expected market shares for new product entries or product modifications are predicted (e.g., Braun et al. 2016; Natter and Feurstein 2002). The quality of all these forecast objectives depends of course on the choice model used for estimation and how good this model can represent consumers' preference structure. The second main area of application is optimal product (line) design. A good overview of this stream and a comparison of the most efficient heuristics developed for this purpose are provided by Belloni et al. (2008). The capability of the estimation model to capture true preferences of consumers (also called parameter recovery) is essential for optimal new product design, as estimated parameters relate to values of product attributes and management is interested to determine optimal attribute levels for their products. It is therefore important to use a choice model for preference estimation that provides a good parameter recovery and/or a good forecasting accuracy depending on related management decisions (like predicting market shares or designing optimal products). The focus of the paper is on the Independence of Irrelevant Alternatives (IIA) property and how good different choice models can handle it as a prerequisite for decision-making.

One of the major limitations of the MNL model is just its IIA property. The IIA property states that the ratio of choice probabilities of two alternatives remains con-

[1] While it is common in some areas to refer to the estimation of choice models based on stated choice data as choice- based conjoint analysis (e.g. in consulting as well as in parts of the academic marketing community), this can be misleading, as discussed in Louviere et al. (2010). The authors have shown that conjoint analysis and discrete choice experiments are fundamentally different, because the former is generally not in line with economic demand theory, while the latter is based on the well-established random utility theory.

stant independent of other available alternatives and hence implies a proportional substitution pattern across alternatives. However, in many situations the IIA property can be unrealistic. The most famous example illustrating this anomaly is the "red-bus/blue-bus paradox" (Debreu 1960). One possible way to capture disproportional substitution patterns are multinomial probit (MNP) models. MNP models are derived under the assumption of normally distributed error terms. However, due to the underlying multidimensional integrals the estimation process is not straightforward and still computationally burdensome (e.g., Ben-Akiva and Lerman 1985; Train 2009). A second approach to capture disproportional substitution patterns are generalized extreme value (GEV) models, which allow for modeling correlations between alternatives. It is a commonality of such models, as the name suggests, that the error terms for all alternatives, i.e. the unobserved parts of utility for all alternatives, are jointly distributed as a GEV. In the simplest case, the distribution represents the product of independent extreme value distributions, which results in a model without correlations between alternatives. Under these conditions, the GEV model degenerates to a MNL model.

The nested multinomial logit (NMNL) model can be considered as an extension of the simple MNL model and is, in addition to the MNL model, the most widely used member of the GEV family. The NMNL assumes that consumers follow a sequential or hierarchical decision making process, that way enabling a partial relaxation of the IIA property. By allowing for different degrees of similarity between subsets of alternatives, specific market structures for modeling consumers' choice decisions can be accommodated. Ben-Akiva (1973) firstly introduced the NMNL model in an application on travel demand. Later, various authors worked on the NMNL model (Ben-Akiva 1979; Daly and Zachary 1978; McFadden 1977, 1978, 1981; Williams 1977). Williams (1977), McFadden (1978) and Daly and Zachary (1978) for example showed that the NMNL model is a special case of the GEV model and is consistent with random utility theory. Since then, the strength of the NMNL model has been recognized so that it found wide application in many areas. Amemiya and Shimono (1989) and Lahiri and Gao (2002) for example analyzed the labor supply choice behavior of elderly people and supposed that some offerings (alternatives) were more similar to each other than to other alternatives. Berkovec and Rust (1985) examined automobile choices of households using the NMNL model. They assumed that customers initially choose one class of cars based on vehicle size and age, and subsequently select a specific brand, type of car or a specific vintage within the chosen class. Falaris (1987) modelled unobserved similarity between states along similar amenities or economic conditions via the NMML model. Although most applications of the NMNL model can be found in the field of transportation and logistics (e.g., Ben-Akiva 1973; Train 1980), the NMNL has also been used in a marketing context. The consideration of different similarities across (subsets of) alternatives is in particular an issue for brand choice modeling (e.g., Anderson and Palma 1992; Baltas et al. 1997; Chintagunta and Vilcassim 1998). On the one hand, brands are often nested with respect to the manufacturers, where products sold by one manufacturer are assumed to be more similar to each other than they are for products sold by different manufacturers (Anderson and Palma 1992). On the other hand, a common choice of nests is the classification of brands into national brands versus private label brands (Baltas et al. 1997) or into subsets of brands with similar prices (Chintagunta and Vilcassim 1998).

It is well-known that accounting for random taste variation in the MNL model, leading to the mixed logit (MXL) model, can strongly soften the IIA property, too. Stated otherwise, while the simple MNL model assumes homogeneous coefficients for all respondents MXL models allow for heterogeneous coefficients across respondents. The first application of the MXL model, also known as random coefficients logit model, dates back to the early 1980s (Boyd and Mellman 1980; Cardell and Dunbar 1980). Since then, the application of MXL models became well-established in marketing research, too (e.g., Allenby and Rossi 1998; Andrews et al. 2002a). It has further been shown that a MXL model can approximate any "true" discrete choice model, given adequate specifications for covariates and the distribution of coefficients (Train 2009). However, if the MXL model is estimated in a frequentist setting, it is not possible to draw inferences on individual-level coefficients, but only on the parameters of the underlying distribution of heterogeneity (e.g., Allenby and Rossi 1998; Rossi and Allenby 2000).

In the mid-1990s, Markov Chain Monte Carlo (MCMC) techniques became available allowing the estimation of the MXL model in a Bayesian setting, too (denoted as HB-MNL model in the following). In contrast to the frequentist MXL approach, the HB-MNL model follows a hierarchical structure that additionally enables the estimation of individual-level coefficients using both information from the heterogeneity distribution of the population (commonly the multivariate normal distribution) and each individual´s choice data (e.g., Allenby et al. 1995; Allenby and Ginter 1995; Lenk et al. 1996). Moore (2004) reported a superior performance of the HB-MNL model as compared to less disaggregated approaches (latent class choice models, fully aggregated choice models) with regard to hit rates in holdout choice tasks, as well as a better or at least comparable performance for the HB-MNL model regarding choice share predictions. Similar results for choice models were provided by Allenby et al. (1998), Huber (1998), Moore et al. (1998), and more recently by Karniouchina et al. (2009). Furthermore, evidence exists that HB models can provide part-worth utilities at the individual respondent level, even when there are insufficient degrees of freedom (Lenk et al. 1996). Today, in order to capture heterogeneity (random taste variation) in discrete choice models, the HB-MNL model represents the state-of-the-art in the marketing literature (Akinc and Vandebroek 2018; Aribarg et al. 2017; Hein et al. 2019a, 2019b; Voleti et al. 2017).

Poirier (1996) proposed a Bayesian framework for estimation, testing, and prediction in the standard NMNL model (i.e., without random taste variation). Following Poirier (1996), Lahiri and Gao (2002) developed an efficient MCMC algorithm for model estimation. At that time, a Bayesian approach for estimating standard NMNL models was particularly welcome since determining the global maximum of a NMNL likelihood function can be extremely cumbersome. Ailawadi et al. (2007) adopted a heterogeneous NMNL (i.e., accounting for random taste variation) to analyze promotion-induced consumer stockpiling in an integrated brand choice / purchase incidence / purchase quantity model. The authors considered the well-known phenomenon that promotions can induce consumers to stockpile multiple items of a promoted brand favoring probably a higher brand consumption, preemptive brand switching, and/or higher repeat purchase probabilities of the brand after the promotion, but they were the first to disentangle these different effects. In particular, the purchase incidence and brand choice parts of their model were treated in the nested logit framework, assuming that a household chooses a specific brand in a considered product category on her/his shopping trip given that the household decided to make a category purchase. As such, the entire set of brands assumed to be available in a product category versus the option to choose none of the items were treated as nests. Model estimation was performed in a frequentist setting using simulated maximum likelihood (Train 2009). Kopalle et al. (2012) studied the impact of considering household heterogeneity in reference price effects on retail brand pricing and category profit implications by means of a NMNL model with two different heterogeneity specifications (finite mixture and HB). The authors used the nested logit framework to model the impact of the reference price both on the purchase incidence of a household to buy versus not to buy in a product category as well as on brand choice given a purchase act. The modeling setup of Kopalle et al. (2012) was therefore similar to Ailawadi et al. (2007) with regard to the nested logit specification, but augmented by a normative modeling step for optimal dynamic product category pricing. Our approach differs from theses two papers in two directions: First, we will compare the statistical performance of a heterogeneous NMNL model to that of the less complex and widespread heterogeneous MNL, estimating both models using Hierarchical Bayes, to analyze the potential benefits from considering correlated errors in addition to heterogeneity. Second, from a modeling perspective, we allow for different nests of „real" alternatives from which an item is chosen, independent of whether an outside or no-purchase option is provided or not, and we further enable different degrees of similarity between subsets of alternatives in different nests.

To the best of our knowledge, there is no paper related to stated choice data that has yet addressed unobserved consumer heterogeneity in the NMNL model using a HB estimation framework. In this paper, we propose a Monte Carlo study using stated choice data in order to analyze the capabilities of the HB-NMNL model under varying data conditions. The advantage of using simulated data is that experimental factors that are assumed to affect the model performance can be varied systematically, and undesirable confounding factors can be held constant. Our expectation is that the HB-NMNL model should outperform the HB-MNL model both in terms of predictive accuracy and goodness of parameter recovery, just because the former explicitly accounts for nested market structures in addition to consumer heterogeneity. On the other hand, Andrews et al. (2002a) found the HB-MNL model seems more flexible than often assumed, as it performed quite robust against violations to its assumption of a unimodal prior distribution on segmented markets. For example, the HB-MNL performed well even if the true preferences followed a mixture of normal distributions (instead of a single normal distribution). The finding that the HB-MNL seems actually to be very robust in multimodal preference settings held true in a recent Monte Carlo study by Goeken et al. (2024), who compared the statistical performance of the HB-MNL not only to a mixture-of-normals MNL, but additionally to a Latent Class MNL and a Dirichlet Process Mixture MNL. Finally, in a different stream of literature on estimating discrete choice models based on aggregate sales data, related to the seminal papers of Berry (1994) and Berry et al. (1995), Grigolon and Verboven (2014) proposed a random coefficients nested logit model and compared it to a random coefficients logit model. While both macro logit models provided comparable price effect predictions for both empirical and simulated data, the random coefficients model tended to underestimate cross-price elasticities and suggested a wider market definition than the random coefficients nested logit model.

Several previous papers compared the statistical performance of competing models using Monte Carlo studies, relating either to traditional conjoint data or stated choice data. Comparisons referred to the performance of different models for metric conjoint data (Andrews et al. 2002b; Vriens et al. 1996), to the performance of the HB-MNL model (e.g. versus the finite mixture MNL model) on segmented markets (Andrews et al. 2002a; Goeken et al. 2024), or to the performance of the HB-MNL model using different levels of information (Wirth 2010), but did not involve the (HB-)NMNL model. In a recent empirical study, Elshiewy et al. (2017) found only marginal differences between the standard NMNL model and the standard MNL model (i.e., without considering respondent heterogeneity), and that the HB-MNL model outperformed both models in terms of goodness-of-fit (as expected). In contrast, Hoffman and Duncan (1988) found empirical differences in part-worth utility structures between NMNL and MNL models (again, both models did not consider random taste variation).

In this paper, we compare the performance of the HB-MNL model to the HB-NMNL model under experimentally varying conditions for stated choice data. On the one hand, the marketing literature has often suggested the standard NMNL model as a remedy for relaxing the IIA property. On the other hand, accounting for consumer heterogeneity by conventional MXL models (random coefficient MNL models) or the HB-MNL model is also known to soften the IIA property to a larger extent. This raises the central research question if an additional relaxation of the IIA property by extending the HB-MNL to additionally accommodate correlations between alternatives has further advantages over the NMNL model. In other words, we want to investigate how robust the HB-MNL model is against violations of the underlying assumptions, especially in the presence of nested structures in data. We will show in Sect. 2.1 in detail that correlated error terms in the HB-NMNL can vary over choice situations, while its heterogeneity component does not, such that our model extension seems particularly promising if time-varying effects on the part of the respondents over choice situations occur. On the other hand, the findings of Grigolon and Verboven (2014) suggest that it can simply be necessary to explicitly accommodate existing sources of market segmentation (like clear nest structures) in a model, which otherwise cannot be fully picked up from the data by addressing heterogeneity alone. To the best of our knowledge, no Monte Carlo study has yet systematically explored the performance of the HB-NMNL model for disaggregate choice data.

Our simulation design leans on the designs of Vriens et al. (1996), Andrews et al. (2002a) and Andrews et al. (2002b). Using statistical criteria for parameter recovery, goodness-of-fit and predictive accuracy we evaluate the comparative performance of the HB-MNL versus the HB-NMNL under varying sample sizes, different nest structures, different levels of respondent heterogeneity, varying number of alternatives within a choice task and different numbers of parameters to be estimated at the individual respondent level.

In the next section, we describe the design of our Monte Carlo study. In particular, we introduce the models, performance measures, as well as the experimental factors and factor levels used. We subsequently present the results of (a) the Monte Carlo study and (b) the empirical study, and discuss implications. Full details on Bayesian model estimation and the data generation process are provided in the Appendix to this paper.

2 Design of the Monte Carlo study

2.1 Model specification and estimation

As explained in the beginning, the IIA assumption is critical in certain situations. If additional information about the similarity of alternatives is provided, it could be useful to partition similar alternatives into subsets, the so-called nests. The ratio of choice probabilities between two alternatives within each nest is independent of the availability of other alternatives so that the IIA property holds within each nest. Accordingly, the ratio of choice probabilities of two alternatives in different nests can depend on other alternatives, which belong to either nests containing these alternatives. While the simple MNL model is fully prone to the IIA property over all alternatives, the standard NMML model only suffers from this property over alternatives within each nest. Stated otherwise, the IIA property does not hold for alternatives in different nests. With the help of NMNL models, such substitution patterns can be handled. NMNL models are obtained by assuming that the vector of error terms has a special type of a GEV distribution. If there are no correlations between alternatives, the distribution degenerates to a product of independent extreme value distributions resulting in the MNL model. For this reason, the NMNL model can be referred to as a generalization of the MNL model. If alternatives can be partitioned into homogeneous subsets, the NMNL model can be applied. Following Train (2009), these subsets are subject to two properties, as mentioned above. On the one hand, the IIA holds within each nest which implies constant cross-elasticities and substitution rates in each nest. On the other hand, cross-elasticities and substitution rates are not constant over alternatives in different nests or, stated otherwise, the ratio of choice probabilities of two alternatives belonging to different nests depends on further alternatives belonging to the same subsets. This characteristic of the NMNL results in the Independence of Irrelevant Nests (IIN) property, which implies that the ratio of choice probabilities of two alternatives belonging to different nests only depends on alternatives in the same nests.

Let's denote the utility that respondent n obtains from alternative j in choice situation s as:

$$U_{njs} = V_{njs} + \varepsilon_{njs}, \tag{1}$$

where $V_{njs} = x_{njs}\beta_n$ represents the deterministic utility and ε_{njs} is the stochastic component. β_n represents the vector of part-worth utilities of respondent n, and x_{njs} is a dummy coded vector for the attribute levels of alternative j in choice situation s

evaluated by respondent n. Assuming that the error terms are jointly distributed as a GEV, the choice probability of alternative j belonging to nest B_k can be written as:

$$\mathrm{P}_{\mathrm{njs}}^{\mathrm{NMNL}}=\frac{\mathrm{e}^{\mathrm{V_{njs}}/\lambda_\mathrm{k}}\left(\sum_{\mathrm{i}\varepsilon\mathrm{B_k}}\mathrm{e}^{\mathrm{V_{nis}}/\lambda_\mathrm{k}}\right)^{\lambda_\mathrm{k}-1}}{\sum_{\mathrm{l}=1}^{\mathrm{K}}\left(\sum_{\mathrm{h}\varepsilon\mathrm{B_l}}\mathrm{e}^{\mathrm{V_{nhs}}/\lambda_\mathrm{l}}\right)^{\lambda_\mathrm{l}}} \tag{2}$$

$B_1, \ldots, B_K$ represent the K disjunct subsets. For two alternatives m and j belonging to the same nest the error terms ε_{nms} and ε_{njs} are correlated and vice versa, for two alternatives belonging to different nests the error terms ε_{nms} and ε_{njs} are uncorrelated. For the latter case, this implies: $\forall j \in B_k, m \in B_l$ with $k \neq l : Cov\,(\varepsilon_{nms}, \varepsilon_{njs}) = 0$. Formula (2) can be used to show that the IIA property does not hold across subsets of alternatives B_k and B_l with $k \neq l$ and that the ratio of probabilities depends on λ_k and λ_l. For two alternatives $\mathrm{j} \in B_k$ and $\mathrm{m} \in B_l$ with $k \neq l$ the ratio of probabilities is:

$$\frac{\mathrm{P}_{\mathrm{njs}}^{\mathrm{NMNL}}}{\mathrm{P}_{\mathrm{nms}}^{\mathrm{NMNL}}}=\frac{\mathrm{e}^{\mathrm{V_{njs}}/\lambda_\mathrm{k}}\left(\sum_{\mathrm{i}\varepsilon\mathrm{B_k}}\mathrm{e}^{\mathrm{V_{nis}}/\lambda_\mathrm{k}}\right)^{\lambda_\mathrm{k}-1}}{\mathrm{e}^{\mathrm{V_{nms}}/\lambda_\mathrm{l}}\left(\sum_{\mathrm{h}\varepsilon\mathrm{B_l}}\mathrm{e}^{\mathrm{V_{nhs}}/\lambda_\mathrm{l}}\right)^{\lambda_\mathrm{l}-1}}. \tag{3}$$

For $k = l$, i.e. alternatives j and m are in the same nest, the sums cancel out and we see that the ratio of probabilities is independent of other alternatives. The coefficients $\lambda_1, \ldots, \lambda_K$ are called log-sum coefficients and measure the degree of independence among the unobserved parts of the utilities of alternatives in the same nest. To be consistent with utility maximization theory, $\lambda_1, \ldots, \lambda_K$ must lie between 0 and 1 (Train 2009). $(1 - \lambda_k) \in [0,1]$ is a measure for the correlation between the stochastic utilities of the alternatives in nest k.[2] For $\lambda_k = 1\ \forall k$ there is no correlation among the stochastic components and the NMNL model simplifies to a MNL model:

$$\mathrm{P}_{njs}^{\mathrm{MNL}}=\frac{e^{V_{njs}}}{\sum_{l=1}^{K}\left(\sum_{h\epsilon B_l}e^{V_{nhs}}\right)}=\frac{e^{V_{njs}}}{\sum_h e^{V_{nhs}}}. \tag{4}$$

If $(1 - \lambda_k)$ increases (or λ_k decreases), the correlation between the stochastic utilities among alternatives in nest k also increases (Ben-Akiva and Lerman 1985). In contrast to the part-worth utilities, the log-sum coefficients are assumed to be fixed across respondents. Alternatively, $\mathrm{P}_{njs}^{\mathrm{NMNL}}$ can also be represented by the product of two MNL models. The first MNL model describes the choice among nests, the second MNL model describes a conditional probability of choosing an alternative given a nest has been chosen. The choice of nests is called upper model and the choice of alternatives within a nest is called lower model (Train 2009).[3] The inclusive value, a logarithmic transformation of the nest specific utilities, links the upper and the lower

[2] Actually, the correlation is a more complex construct, but McFadden (1978) used 1-λ_k as an index of correlation or independence.

[3] Note that Ailawadi et al. (2007) and Kopalle et al. (2012) used a binary logit as upper model to represent purchase incidence of a consumer (i.e. whether a consumer buys in a product category or not), while

models and (multiplied by λ_k) is a measure for the expected utility that respondent n receives from an alternative belonging to nest B_k.

In order to motivate why extending the HB-MNL to the HB-NMNL by adding correlated error terms might be beneficial (i.e. can lead to an additional relaxation of the IIA property), it is reasonable to show, from a conceptual perspective, that the error term ε_{njs} of the nested logit model can be divided into two parts, leaning on Grigolon and Verboven (2014, pp. 918) and following Berry (1994) and Cardell (1997).[4] Remember that each alternative j is assigned to exactly one nest, then:

$$\varepsilon_{njs} = \zeta_{nks} + \lambda\widetilde{\varepsilon}_{njs}, \tag{5}$$

where $\widetilde{\varepsilon}_{njs}$ is the usual i.i.d. extreme value distributed error, λ is the (global) log-sum coefficient that describes the average substitutability between alternatives in a nest, and ζ_{nks} represents a nest-specific utility shock related to all alternatives in nest *k*. The existence of nest-specific utility shocks leads to correlated errors.

Replacing now the error term ε_{njs} of the utility function (1) by Eq. (5) and indicating the membership of alternative *j* to nest *k* via the dummy d_{jks} shows how the two components *random taste variation* and *correlated errors* affect a respondent's utility:

$$U_{njs} = x_{njs}\beta_n + \sum\nolimits_{k=1}^{K} d_{jks}\zeta_{nks} + \lambda\widetilde{\varepsilon}_{njs}. \tag{6}$$

Eq. (6) reveals that taste heterogeneity (first term) is constant across the *S* choice situations, while correlated errors can vary across choice situations. Even if correlations between part-worth utilities β_n were accommodated by estimating a full covariance matrix, as is nowadays standard in modeling stated individual choices using an HB framework (and also as in our study), ζ_{nks} might help to capture further aspects from the data and improve the model performance.

In the simplest case, the additional error term might compensate for a violation of the assumed normality of the individual part-worths β_n. However, since the heterogeneity component (heterogeneity of respondents in part-worth utility structures) is constant in time in the HB-NMNL, only the correlated error terms can pick up potential time-varying shocks or effects over choice occasions. For example, although respondents will be typically informed prior to a discrete choice experiment about product attributes and levels used to describe alternatives in the repeated choice situations, the information processing of respondents might change over choice situations due to learning effects (e.g. Day et al. 2012; Hess et al. 2012). This means that respondents get to know new product configurations in the choice tasks and then might adjust their answers accordingly over the course of the experiment. Also, respondents might become disengaged due to fatigue or boredom as the number of choice tasks

we in principle allow for an arbitrary number of different nests of alternatives from which respondents choose an item in repeated choice situations, including the none option as additional nest if considered.

[4] We thank one of the anonymous reviewers for pointing us to the paper by Grigolon and Verboven (2014), which helped us a lot to improve the motivation and positioning of our modeling approach.

increase (e.g. Day et al. 2012; Savage and Waldman 2008). Both learning and disengagement effects but also overwhelm effects as a result of too complex choice situations (Huber et al. 2018) can lead to simplification strategies of respondents across choice situations in empirical settings (e.g. Hein et al. 2019a). While learning effects decrease the error variance over the course of the choice experiment, disengagement or overwhelm effects increase the error variance (e.g. Boxebeld 2024). However, for this reason, simplification strategies might not be the best argument for justifying time-varying utility shocks, since the related processes typically manifest as gradual changes in scale (an increasing or decreasing error variance) rather than as discrete shocks to the mean utility.[5]

In labeled choice experiments where each alternative is shown in each choice situation, changes in the position of the alternatives over choice situations can lead to time-varying correlations. In contrast to simplification strategies, design-induced position effects offer a stronger rationale for the existence of correlated errors across choice sets. Position effects arise from the order of presentation of the alternatives within a choice set, and there is strong empirical evidence that they can affect respondent preferences, see the systematic literature review about the position effect (also referred to as 'alternative ordering effect') provided by Boxebeld (2024). The study by Boxebeld (2024) identified ten different empirical papers on this topic, conducted between 1994 and 2022, of which seven reported significant position effects. In addition, different versions of choice task questionnaires are often used to increase the efficiency of choice task designs, inducing potentially different individual time-varying effects over choice situations. Such changes in the presentation order of the alternatives over the sequence of choice sets are similar to changes in shelf placements of products frequently observed in real choice settings (e.g. in supermarkets). Beyond the position effect, a number of studies have further analyzed the choice set ordering effect. i.e. whether the position of a choice set (or several choice sets) in the choice task design impacts respondent preferences and error variances (e.g. Day and Prades 2010, Carlsson et al. 2012; Day et al. 2012, Nguyen et al. 2021, as well as again Boxebeld 2024 for an overview of further studies in this field). The choice set ordering effect was also found to be significant in many studies, but importantly, as opposed to the position effect which is strongly associated with lexicographic consumer behavior, the choice set ordering effect seems to be driven by learning and fatigue, or anchoring effects (e.g. Boxebeld 2024).

Furthermore, labeled choice experiments strongly favor the formation or existence of different nests on the part of the respondents (e.g. low-priced low-quality versus high-priced high quality brands), and hence favor the application of a nested logit model framework. In practice, it usually depends on the underlying research question whether an experiment is labeled or not. Pricing studies almost always require that all brands (alternatives) are available in each choice task, otherwise the proportion of the outside good becomes too large. Of course, not all choice experiments are labeled, for example in case of feature discrete choice experiments (feature conjoints) conducted to reveal preferences for new non-price features of products. Here, it is certainly con-

[5] We thank an anonymous reviewer for pointing this out.

ceivable not always to show all alternatives or to omit brand alternatives entirely.[6] In our Monte Carlo study, we will use a labeled choice experiment with varying overall utilities of alternatives over choice situations. Importantly, Grigolon and Verboven (2014, p. 917) concluded from their study *„that it can be important to account for sources of market segmentation“,* which in case of the existence of different nests („segments“) might not be fully captured by the continuous heterogeneity distribution for the part-worths alone but by the additional nest-specific utility shocks ζ_{nks}, to improve a model‘s performance. Note that in this situation the nest-specific utility shocks need not necessarily be time-varying.

In order to obtain part-worth utilities at the individual respondent level both the MNL and the NMNL models are estimated applying Bayesian procedures. MCMC methods are used to generate draws from posterior distributions. Specifically Gibbs Sampling, first introduced by Geman and Geman (1984), together with the Metropolis-Hastings algorithm, introduced by Metropolis et al. (1953) and Hastings (1970), are applied. For more general explanations of Gibbs sampling and the Metropolis-Hastings algorithm see e.g. Gelman et al. (2014). The estimation of MXL models using a Bayesian framework, also known as hierarchical Bayes, is described in detail in Train (2009) and Train and Sonnier (2005). The inclusion of fixed effects in the estimation process of MXL models is described in e.g. Train (2001). The estimation of our heterogeneous NMNL model using a Bayesian framework is described in detail in the Appendix to this paper, see Appendix 1. Following Train (2001), the addition of the fixed effects (here: the log-sum coefficients $\lambda_1, \ldots, \lambda_K$) requires a fourth layer in the Gibbs sampling. Drawing the fixed coefficients simultaneously with the individual part-worth utilities would cause the fixed effects to be random across respondents, which is inconsistent with the assumption of fixed effects. Therefore, the estimation of fixed effects needs to be separated from that of random effects. Accordingly, the log-sum coefficients are determined separately in the fourth layer of the model conditional on the current values of the random coefficients, and based only on the likelihood function (there is no need to specify an additional prior for the fixed effects).

Model estimation is carried out using the publicly available *R* software package *RSGHB,* authored by Dumont et al. (2019). The software provides functions for estimating models using Hierarchical Bayes, allowing (a) a direct specification of the likelihood function, (b) flexible specifications of model parameters as either random effects with continuous distributions or fixed effects (the latter we use for the log-sum coefficients), and (c) the choice of various parameter distributions. A part of the functions in this package is based on the codes of Kenneth Train for Hierarchical Bayesian estimation, and Stephane Hess and the Choice Modeling Centre also contributed to the package. Details about *RSGHB* can be found at https://doi.org/10.32614/CRAN.package.RSGHB.[7]

[6] According to Peter Kurz, formerly Head of Research and Development at Kantar TNS (2001-2018) and one of the world´s leading practitioners for discrete choice models, stated choice experiments were seldom unlabeled under his leadership. Of course, this may also depend on the specialization of the market research institute.

[7] As an alternative to *R*, the programming language *Stan* also enables statistical modeling using Bayesian inference. Here, Van Horn and McDonnel Feit (2017) have provided a publicly available *Stan* package

2.2 Experimental factors and data generation

The choice of the experimental factors and factor levels for our Monte Carlo study leans on the experimental designs of Vriens et al. (1996), Andrews et al. (2002a) and Andrews et al. (2002b). Of course, in order to represent a realistic nested market structure, the experimental factors need to be adjusted adequately. In the following, the nested structure is accommodated considering different price and quality tiers that are assumed to influence the purchasing behavior of consumers (e.g., Blattberg and Wisniewski 1989; Sivakumar 1995). Gabor and Granger (1966) already pointed out that price can be an indicator of quality instead of only representing the opportunity cost of purchasing (also see Chatterjee and Kumar (2017) or Rao and Monroe (1989)). Further, Blattberg and Wisniewski (1989) showed that brands with a higher price and a higher quality in particular steal market shares from similar brands in the same price-quality tier. Analogously, brands belonging to a lower price-quality tier tend to affect, almost exclusively, the market share of brands in their own price-quality tier, too. Sivakumar (1995) stated that this pattern of competition leads to a two-stage purchase decision process. In the first stage, the respondent selects a price-quality tier, in the second stage a corresponding brand. Consequently, different degrees of similarities between alternatives exist in this market structure suggesting the application of the NMNL model to handle the resulting IIA problems (Allenby and Rossi 1991), that probably cannot be solved by addressing heterogeneity alone. For our Monte Carlo study, we adopt this kind of market structure and specify two nests of alternatives representing two different price-quality tiers, one representing higher quality and more expensive brands, and one containing less expensive alternatives at lower quality. Consequently, the data generation process needs to consider that parts of respondents tend to prefer higher quality brands, while other respondents are rather price-sensitive and therefore tend to lower quality brands. Nevertheless, based on empirical evidence, tier switching is made possible in our simulation study in a way that consumers may switch up from lower to higher quality brands or the reverse under certain price-quality constellations (e.g. Allenby and Rossi 1991; Blattberg et al. 1995), see Appendix 2 for details of the data generation process. Of course, more granular nested market structures (e.g. a low, mid, and high quality tier structure) are possible, which in practice strongly depends on the product category. To hold the number of treatments in our Monte Carlo study manageable, however, we concentrate in the following on a two-submarket structure, which is often observed in real markets. Still, we discuss this issue in detail at the beginning of Sect. 4 (empirical study).

Overall, five factors were experimentally manipulated in the current study: (1) the number of respondents per nest (i.e. respondents tending to higher versus lower quality brands), (2) the number of alternatives (brands) offered per nest, (3) the number of parameters to be estimated at the individual respondent level (in addition to the

for estimating a hierarchical nested multinomial logit model, that allows in its current implementation to model purchase incidence and brand choice (choose none versus an actual alternative), like in Ailawadi et al. (2007) and Kopalle et al. (2012). It is certainly straightforward to extend the package to the more general case of more nests in Stan.

brand constants for the alternatives), (4) the degree of similarity between alternatives, and (5) the level of preference heterogeneity. Factors (2) and (4) determine the nest structure, where high (low) quality of alternatives is reflected by larger (smaller) brand constants and the similarity between alternatives is specified by the log-sum coefficients. Factor (3) defines the model complexity which depends on the number of attributes (5 vs. 9) and attribute levels (3 vs. 5) specified in addition to the brand constants. The experimental factors and factor levels included in the Monte Carlo study are summarized in Table 1.

In total, there are $5 \times 2^2 \times 4^2 = 320$ experimental data conditions (treatments), and with one replication and two models (HB-MNL vs. HB-NMNL) to be estimated under each treatment we obtain $320 \times 2 \times 2 = 1{,}280$ observations for statistical analysis. In empirical applications, only the alternatives (brands), the number of parameters to be estimated at the individual level, and the total sample size (n = 1000 in our case) are observable prior to estimation. In contrast, the partitioning of the respondents reflecting their general preference either towards higher-quality higher-priced respectively lower-quality lower-priced brands, the log-sum coefficients, as well as the degree of heterogeneity are unobservable in empirical data prior to estimation. Note that, as mentioned above, we will generate preferences that fit these nesting structures (i.e. specifying the number of respondents tending to higher- vs. lower-quality brands, generating a higher importance of the brand attribute for respondents tending to higher-quality brands and vice versa, etc.), but we also allow for quality tier competition. If switching between the price-quality tiers would not be possible,

Table 1 Experimental factors included in the study

Factor	Factor levels	# Factor levels	Short-cut[b]
1. Nest size (number of respondents with a general tendency in their preference towards a nest)	$N_1 = 125, N_2 = 875$	5	125 875
	$N_1 = 250, N_2 = 750$		250 750
	$N_1 = 500, N_2 = 500$		500 500
	$N_1 = 750, N_2 = 250$		750 250
	$N_1 = 875, N_2 = 125$		875 125
2. Number of alternatives per nest (brands)	2	2	2
	3		3
3. Model complexity (number of parameters excluding brand constants)[a]	10	4	A5L3[c]
	18		A9L3
	20		A5L5
	36		A9L5
4. Similarity between alternatives (log-sum coefficients)	$\lambda_1 = \lambda_2 = 0.2$	4	0202
	$\lambda_1 = \lambda_2 = 0.8$		0808
	$\lambda_1 = 0.2, \lambda_2 = 0.8$		0208
	$\lambda_1 = 0.8, \lambda_2 = 0.2$		0802
5. Preference heterogeneity	Low	2	005
	High		025

[a]Note that for L categorically coded attributes with M levels each, L times (M-1) parameters need to be estimated independent whether a dummy or effects coding is used

[b]These shortcuts are used later for the presentation of results

[c]A5L3: 5 attributes à 3 levels (analogously for the other shortcuts)

the nested logit model would degenerate to a model with fixed choice set heterogeneity (e.g. Crawford et al. 2021).

Since HB models borrow information from all individuals (respondents) for parameter estimation, the degree of random taste variation in a sample might affect the individual-level estimates. The experimental factor heterogeneity with its two levels (factor 5) accounts for the amount of preference heterogeneity in the data generation process and hence enables to consider more or less random taste variation in the part-worth structures across respondents (also compare Andrews et al. (2002a), Hein et al. (2019b), Vriens et al. (1996) and Andrews et al. (2002b)). We expect that a smaller level of heterogeneity should have a positive effect on the statistical model performance, at least should make it easier to recover "true" parameter values (e.g., Andrews et al. 2002a, 2002b; Vriens et al. 1996). Furthermore, we conjecture that it should be more difficult to identify a nested structure when preferences across respondents are more heterogeneous. For this reason, we expect that the HB-NMNL model does not achieve substantial advantages over the HB-MNL model in treatments with high sample heterogeneity. We varied the amount of heterogeneity according to Andrews et al. (2002a) by setting once the variance of part-worths across respondents to a low level (0.05) and once to a high level (0.25).[8]

Following previous empirical studies, we assume that alternatives are partitioned into two nests, one consisting of high-quality brands and one consisting of low-quality brands (e.g., Silberhorn et al. 2008; Sivakumar 1995). Sivakumar (1995) distinguished a high-quality cluster and a low-quality cluster along the dimensions price and quality. In our study, quality is associated with the brand attribute and quantified by the value of the brand constant. The price is set conditional on the quality of the brand, see the data generation process described in Appendix 2 for details. For our choice task design we assume that each brand appears once in each choice set, so that the number of alternatives per choice set is set to 4 or 6, depending if either 2 or 3 brands define the high-quality and low-quality nests, respectively (factor 2).

By allowing for different degrees of similarity between subsets of alternatives different substitution patterns can be modelled. The similarity between alternatives belonging to the same nest is generated by the log-sum coefficient (factor 4), which is a measure for the degree of independence of the unobserved part of utility among alternatives in the same nest. As Train (2009) pointed out, $1 - \lambda_k$ can be used as a measure of correlation. If λ_k increases, the correlation between alternatives decreases. We specify four different levels of similarity: (a) high similarity between alternatives in both nests($\lambda_1 = \lambda_2 = 0.2$), (b) small similarity between alternatives in both nests($\lambda_1 = \lambda_2 = 0.8$), (c) high similarity between alternatives in the first nest and small similarity between alternatives in the second nest $(\lambda_1 = 0.2, \lambda_2 = 0.8)$ and (d) small similarity between alternatives in the first nest and high similarity between alternatives in the second nest($\lambda_1 = 0.8, \lambda_2 = 0.2$). The factor levels for the log-sum parameters are chosen based on a Monte Carlo study proposed by Garrow et

[8] Andrews et al. (2002a) increased the higher end variance for the heterogeneity distribution in comparison to Andrews et al. (2002b) and Vriens et al. (1996) from 0.1 to 0.25 to ensure a profound understanding of the effects of consumer heterogeneity. Previous studies have shown that the former variation between 0.05 and 0.1 was too weak to provoke noticeable changes in some performance measures.

al. (2010), who found the values of 0.8 and 0.2 to correspond with the minimum and maximum amount of correlation that existed in their NMNL structure. Since the MNL model assumes independent error terms, we expect that a high similarity of alternatives in at least one of the nests should favor the NMNL model. In the case of low correlations in both nests, we instead expect a similar performance of the MNL and NMNL models.

A larger number of individual parameters in the model (factor 3) given a certain amount of information at the individual respondent level (as provided by the number of choices per respondent) reduces the number of degrees of freedom for estimation. As a consequence, a larger number of parameters at an individual level should yield less reliable parameter estimates independent of the type of model (HB-MNL or HB-NMNL). Note that the number of brands per nest (factor 2) also affects the number of parameters in the model.

The partitioning of respondents (factor 1) refers to the number of respondents with a general preference for higher-priced higher-quality versus lower-priced lower-quality brands. In the simplest case, the 1,000 respondents are evenly split in their preferences towards either nest/tier (500/500). In the asymmetric cases, $1/4$ or $1/8$ of the respondents tend to prefer the alternatives in one tier, while $3/4$ or $7/8$ of the respondents the alternatives in the other tier, leading to the four combinations 250/750, 125/875, 750/250, and 875/125. Without loss of generality, we assume that the first nest represents alternatives with a higher quality and therefore consists of more expensive brands, and that the second nest contains less expensive alternatives of lower quality. As will be explained in more detail in Appendix 2 (details on the data generation process), we defined nest-specific utility ranges for attributes and in particular for the two characteristics reflecting the brand and price attribute to define two different price-quality tiers. Therefore, although the factor levels 250/750 and 750/250 respectively 125/875 and 875/125 represent equal conditions regarding the partitioning of respondents, they represent different conditions with respect to nest-specific utility patterns: the smaller (larger) nest once represents high quality and high priced brands and once low quality and low priced brands. By manipulating factor 1 we want to explore the capabilities of the HB-NMNL and the HB-MNL models to recover different nest sizes with different nest structures (here representing different price-quality tiers). Stated more formally, taking into account the hierarchical structure of the Bayesian estimation framework, the underlying question is how well the posterior distribution can uncover multimodal and correlated utility structures in the presence of sparse data (small nest sizes), although the prior distribution of both models (MNL, NMNL) is unimodal and tends to shrink units toward the population mean. By studying histograms of the posterior parameter estimates, Andrews et al. (2002a) showed for the HB-MNL model that the posterior distribution could turn out multimodal even when the prior distribution is unimodal. Furthermore, as the total number of respondents is fixed to 1,000, i.e. the population information does not vary, no further effects on individual part-worth utilities can be expected. The total number of respondents used in this study is chosen somewhat larger than in previous Monte Carlo studies to account for sample sizes common in practical studies, but is also in line with the trend of an increasing number of artificial respondents in more recent Monte Carlo studies. The sequence of magnitudes of nest specific sample sizes still leans on the studies of Andrews et al. (2002a), Vriens et al. (1996), Andrews et al. (2002b) and Wirth (2010).[9] We do not expect that a variation of the total number of respondents favors either the MNL or the NMNL model.

In the following, we provide the basic information how the synthetic data sets were generated for our Monte Carlo study. For a detailed description of the data generation process, we refer the reader to Appendix 2 of this paper. The data generation process can be divided into the construction of the experimental design, the generation of individual part-worth utilities, and the generation of choice decisions given the generated design and the individual part-worth utilities. The data generation process is illustrated in Fig. 1.

Design: Based on the number of brands per nest (factor 2) and the number of non-brand attributes and attribute levels (factor 3), D-efficient choice designs were created by using computer-generated design algorithms implemented in SAS (Kuhfeld et al. 1994; Kuhfeld 1997; Zwerina et al. 1996). The number-of-levels effect was controlled by using (near) symmetrical designs (Verlegh et al. 2002), i.e. all non-brand attributes had either three or five levels. As already mentioned, we assumed that each brand appears in each choice set so that the number of alternatives per choice set was 4 or 6. Thus, choice designs had four or six alternatives and five or nine attributes per alternative. Without loss of generality, we did not include an outside good in our

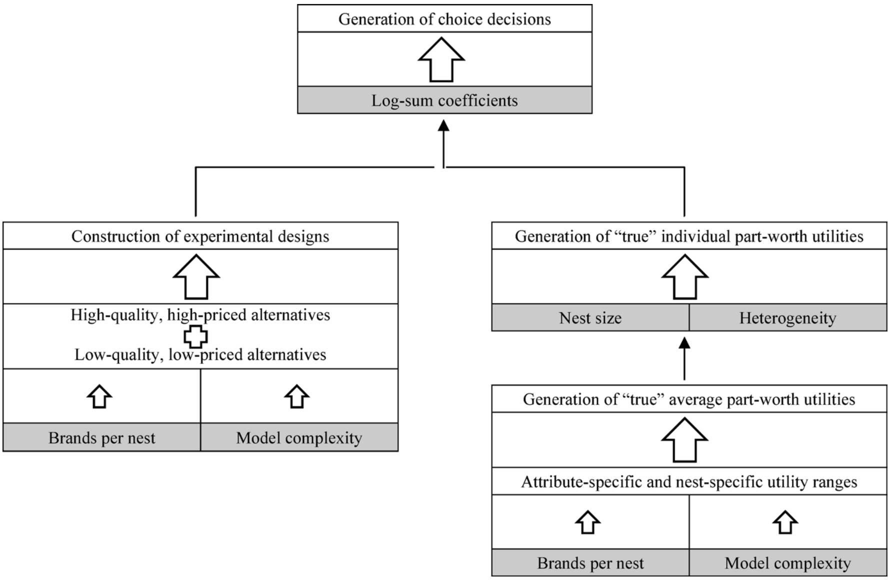


Fig. 1 Data generation process. The experimental factors considered at the different stages of the data generation process are highlighted in grey

[9] Vriens et al. (1996) varied the number of simulated respondents at two levels (100 and 200) and defined each segment (2 vs. 4) to consist of an equal number of respondents. Andrews et al. (2002b) generated 150 respondents, who were randomly assigned to 2 or 3 segments. Andrews et al. (2002a) used a larger number of subjects (75, 200, 400). The corresponding segment sizes for 2 or 3 segments were generated randomly. Wirth (2010) generated 150, 300 or 600 respondents, but did not consider segment-specific part-worth structures.

Monte Carlo study, just because we consider labeled choice experiments where all brands of the two nests, i.e. all inside goods, were always available in each choice set (albeit with varying non-brand attribute levels and therefore varying overall utilities). Including an outside good in this case would not have provided any additional insights regarding the statistical assessment of the two models. Rather, adding a no-purchase option would have required the specification of none utilities or even to include the no-purchase option as an additional experimental factor. However, as is often the situation in empirical studies, if not all inside goods can be included in each choice set, and/or if related quantitites like price elasticities or willingness-to-pay are of primary interest, and/or in studies using market data in marketing (e.g. Chib et al. 2004; Ailawadi et al. 2007; Kopalle et al. 2012), considering an outside good "as additional nest" is mandatory to prevent biased implications.

Following two meta-analyses of commercial stated choice modelling studies conducted by Hoogerbrugge and van der Wagt (2006) and Kurz and Binner (2012), we further set the number of choice tasks per respondent to 15. According to Blattberg and Wisniewski (1989) and Sivakumar (1995), brand (factor 2) and price are appropriate attributes to distinguish between high- and low-quality brand tiers. Therefore, two out of four alternatives or three out of six alternatives in our study (depending on the setting of factor 2) were defined as high-quality brands, the remaining two or three brands as low-quality brands. To generate related price tiers, one of the non-brand attributes was fixed as a hypothetical price attribute with three or five levels. In order to obtain a nested structure, the final design was generated in such a way that higher prices tended to occur together with high-quality brands, and lower prices were assigned to low-quality products. These specific price-quality patterns were accomplished by defining adequate occurrence probabilities (see Appendix 2 for details).

Part-worth utilities: Based on the generated efficient designs comprising 15 choice tasks per respondent, the next step in the data generation process was to simulate choices of 1,000 artificial respondents. Therefore, "true" individual-level part-worth utilities for each respondent had to be generated, and for each of the 640 data sets (320 treatments with one replication; compare Table 1) part-worth utility structures were each time randomly generated. In a first step, we leaned on the Monte Carlo study of Vriens et al. (1996) who used six attributes with three levels each and generated "true" mean part-worth utilities of respondents (mean vector b_{TRUE}) for each attribute according to a uniform distribution from the range between -1.7 to $+1.7$, resulting in a maximum utility range of -10.2 to $+10.2$ summed across all six attributes. However, when assuming a nested structure as in our study, nest-specific utility ranges need to be generated. We therefore assumed that the brand attribute has a mean importance of 50% for respondents with a general preference for higher-quality brands, while a mean importance of only 20% for respondents with a general preference for lower-quality brands. Conversely, we set the mean importance of the price attribute to 50% for respondents tending to the low-quality nest, and to only 20% for respondents tending to the high-quality nest. The remaining 30% importance were split evenly between the remaining four or eight attributes (compare Table 1). Therefore, the brand (as proxy for quality) receives much more importance than the price in the high-quality tier, while the price plays a much larger role than the brand in the low-quality tier. Stated otherwise, different from Vriens et al. (1996) who considered each of the six attributes to be equally important, the maximum utility range of -10.2 to $+10.2$ was scaled depending on nest-specific differences in importance for brand and price. All details regarding the assignment of the generated (mean) brand constants and (mean) price part-worths to the high- and low-quality nests, the generation of individual part-worth utilities by adding a heterogeneity component to the generated mean part-worths, as well as the maintenance of order constraints for both mean and individual price part-worths can be found in Appendix 2. As a result of this specific data generation process, we obtained one group of respondents (N_1) who were less price-sensitive and prefer alternatives of higher quality (brand is more important than price), and a second group of respondents (N_2) who are more price-sensitive and are satisfied with alternatives of lower quality (price is more important than brand). To get this specific multimodal and nested structure was the main goal of the data generation process.

Simulation of choices: Based on the generated choice task designs and the generated "true" individual part-worth utilities, deterministic utilities $V_{njs} = \beta_n' x_{njs}$ for each alternative in each of the 15 choice sets could be computed for each respondent. Different from simulating choices for respondents under a MNL model, simulating choices based on error components that replicate the assumed correlation of a nested logit structure is not such straightforward (as error terms are identically but not independently distributed). Fortunately, Garrow et al. (2010) compared three methods to determine simulated choices of respondents for the NMNL model, with a clear recommendation in favor of "roulette wheel selection" (Lipowski and Lipowska 2012), that uses the closed-form nested logit probabilities according to Eq. (2). In short, the choice of an alternative is proportional to its nested logit probability $\mathrm{P}_{\mathrm{njs}}^{\mathrm{NMNL}}$ obtained from Eq. (2), which can be formally accomplished by standard uniform draws. Based on the simulated choices, the individual-level part-worths β_n as well as the population-level parameters b (mean part-worths), the covariance matrix W, and in case of the NMNL model the log-sum coefficients λ were re-estimated using the HB-NMNL and HB-MNL models.

Since the data generation process provides more or less distinct nested logit structures, the statistical performance of the HB-MNL regarding parameter recovery, model fit and predictive validity is of particular interest (as compared to the HB-NMNL model).

2.3 Measures of performance

A total of 15,000 simulated choices (15 choice sets per respondent, 1,000 respondents) were collected for each of the 640 data sets considered (320 treatments, one replication). For each data set, we randomly split the 15,000 observations into an estimation and validation sample, where 75% of the choices were used for model estimation and 25% of the choices for model validation (i.e., to assess the predictive accuracy). After estimating the two models (HB-NMNL, HB-MNL) for each treatment and replication, the resulting 1,280 observations (640 data sets times two models) were then subjected to analysis of variance, i.e. the type of model is used as additional factor in the ANOVAs. Following, Andrews et al. (2002a), Vriens et al. (1996) and Andrews et al. (2002b), the statistical model performance is evaluated based on several measures for parameter recovery, goodness-of-fit, and predictive accuracy. In particular, we used three measures for parameter recovery, four measures for goodness-of-fit, and five measures for predictive validity. The ANOVAs were based on the part-worth means of the posterior distributions (point estimates).

Parameter recovery was measured by the mean Pearson correlation between the generated ("true") and the re-estimated individual part-worth utilities. Since Pearson correlations are not interval-scaled, the individual-level correlations were at first rescaled using Fisher's z-transformation. After averaging the rescaled values the mean Pearson correlation was obtained by the equivalent retransformation of the rescaled averages. As further measures of parameter recovery, we calculated the root mean square error (RMSE) and the mean absolute error (MAE) between the "true" part-worth utilities (β_{nal}) and the re-estimated part-worth utilities $(\widehat{\beta}_{nal})$:

$$\begin{aligned} RMSE\left(\beta\right) &= \sqrt{\frac{\sum_{n=1}^{N}\sum_{a=1}^{A}\sum_{l=1}^{L_a}\left(\hat{\beta}_{nal}-\beta_{nal}\right)^2}{N\sum_{a=1}^{A}L_a}}, \\ MAE\left(\beta\right) &= \frac{\sum_{n=1}^{N}\sum_{a=1}^{A}\sum_{l=1}^{L_a}\left|\hat{\beta}_{nal}-\beta_{nal}\right|}{N\sum_{a=1}^{A}L_a}. \end{aligned} \tag{7}$$

N denotes the total number of respondents, A the number of attributes and L_a the number of attribute levels per attribute (factors 2 and 3). While the MAE measures the average absolute deviation between the "true" and re-estimated individual part-worths, the RMSE weights large deviations more strongly than small deviations. Since an accurate parameter recovery is crucial for product (line) design or pricing decisions, a stronger penalization of larger deviations might be favorable from a managerial point of view.

To assess model fit the log-likelihood (LL), the percent certainty (PC), the Brier score (BS), and the Spherical score (SpS) were used. LL, BS, and SpS were corrected by the total number of choice tasks and in addition by the number of alternatives within choice tasks (BS) to enable a comparison across treatments that differ in those parameters (also compare Hein et al. (2019a), Andrews et al. (2002a)). Note that the number of choice tasks per respondent can differ across respondents in the estimation sample due to the random split of the data into estimation (75%) and validation (25%) sample. Accordingly, the LL is calculated by

$$\ln\left(L\left(\widehat{\beta}\right)\right) = \frac{\sum_{n=1}^{N}\sum_{c=1}^{C_n}\sum_{j=1}^{J}Y_{njc}\ln(\widehat{P}_{njc})}{\sum_{n=1}^{N}C_n}, \tag{8}$$

where C_n denotes the number of choice sets for respondent n in the estimation sample. The number of alternatives J in a choice set depends on the number of brands per nest (factor 2). Y_{njc} is an indicator variable for the chosen alternative j in choice set

c, and $\widehat{P}_{njc}$ is the estimated probability that respondent n chooses alternative j from choice set c. While the LL only considers the probabilities of alternatives that were chosen by respondents from their choice sets $(Y_{njc} = 1)$, the BS (Brier 1950; Gneiting and Raftery 2007) and the SpS (e.g., Gneiting and Raftery 2007) in addition use the probabilities of all non-chosen alternatives $(Y_{njc} = 0)$ and therefore exploit the information of the complete predictive distribution of choice probabilities:

$$
\begin{aligned}
Brier\ score\left(\hat{\beta}\right) &= -\frac{\sum_{n=1}^{N}\sum_{c=1}^{C_n}\sum_{j=1}^{J}(Y_{njc}-\hat{P}_{njc})^2}{\sum_{n=1}^{N}C_nJ} \\
Spherical\ score\left(\hat{\beta}\right) &= \frac{\sum_{n=1}^{N}\sum_{c=1}^{C_n}\left(\hat{P}_{nj^*c}/\sqrt{\sum_{j=1}^{J}\left(\hat{P}_{njc}\right)^2}\right)}{\sum_{n=1}^{N}C_n},
\end{aligned}
\tag{9}
$$

where j^* denotes the alternative chosen by respondent n from choice set c. The percent certainty compares the LL of the estimated "full" model (model with covariates) to the LL of the null model (model without any covariates) (Hauser 1978; Ogawa 1987):

$$
PC = \frac{LL_{full} - LL_{null}}{-LL_{null}}, \tag{10}
$$

where LL_{full} and LL_{null} denote the LL of the estimated model and the null LL, respectively.

Predictive accuracy was measured by the hit rate, the root mean square error between "true" and predicted deterministic utilities (RMSE(V)), the LL, the BS and the SpS. The hit rate measures the percentage of first choice hits across respondents in the validation sample (Andrews et al. 2002b; Vriens et al. 1996). A first choice hit is counted if a respondent's simulated choice behavior was predicted correctly, i.e. if the alternative chosen by a respondent from a choice set has been assigned the highest predicted deterministic utility among all alternatives offered in that choice set (Andrews et al. 2002b; Elrod and Kumar 1989). The RMSE(V) between "true" and predicted deterministic utilities is calculated by (Andrews et al. 2002b; Paetz and Steiner 2017):

$$
RMSE\,(V) = \sqrt{\frac{\sum_{n=1}^{N}\sum_{c=1}^{C_n^*}\sum_{j=1}^{J}\left(\widehat{V}_{njc} - V_{njc}\right)^2}{\sum_{n=1}^{N}C_n^*J}}. \tag{11}
$$

Due to the random split of the data, the number of choice tasks per respondent can also differ across respondents in the validation sample (denoted as C_n^*). Like for measuring model fit, the LL, the BS and the SpS can be calculated for the validation sample, too (e.g., Abramson et al. 2000).

3 Results of the Monte Carlo study

The impact of the five experimental factors and the two models (HB-NMNL vs. HB-MNL) on each of the twelve measures of performance was examined by analysis of variance for main effects and first-order interaction effects. The ANOVAs were based on a total of 1,280 observations with 1,200 degrees of freedom for error. Table 6 in the Appendix to this paper (Appendix 3: Additional Tables) displays F-Tests for main and interaction effects and corresponding p-values as well as explained variances of the ANOVAs for each performance measure. R-squares (adjusted R-squares) range between 0.284 (0.236) and 0.568 (0.540), whereas most values are greater than 0.4. Table 6 shows that most of the main effects (93%) are significant ($p<0.05$), indicating differences in the measures of performance between the particular factor levels. Interestingly, the HB-NMNL and the HB-MNL models do not differ significantly with regard to the hit rate, the BS and the SpS in the validation sample ($p>0.1$), i.e., the two models predict similarly accurate across treatments. The highest F-values for differences between the two models can be observed for parameter recovery measured in absolute terms (RMSE: 141; MAE: 182). Furthermore, both the partitioning of respondents regarding the nest-specific utility structures (factor 1) and the number of brands per nest (factor 2) do not significantly affect the mean correlation. High F-values of the number of brands per nest and the degree of similarity between alternatives for nearly all model fit measures, as well as high F-values of the number of brands and the sample heterogeneity for nearly all predictive performance measures point to substantial differences between factor levels of these factors with regard to the model performance. Further, 50% of the first-order interaction effects turn out significant. Here, the high F-values for the interaction between the degree of similarity between alternatives and the type of model on parameter recovery measured in terms of RMSE and MAE are noticeable. In addition, nearly all interaction effects between the type of model and any of the other five experimental factors with regard to the predictive validity (23 out of 25 interactions) turn out not significant ($p>0.05$).

However, since even small differences between factor levels may become significant due to the large sample size ($N=1{,}280$), we further computed related effect sizes. Table 2 displays Eta squares (η^2) which we interpret using the guidelines of Cohen (1988) ($\eta^2=0.01$ small effect, $\eta^2=0.06$ medium effect, $\eta^2=0.14$ large effect). In the following, we concentrate on the interpretation of those factors which show at least medium to large effect sizes. The by far largest effect sizes can be observed for factor 4 (similarity between alternatives) with very large effect sizes for all goodness-of-fit measures (>0.26), and medium effect sizes for RMSE and MAE as measures of parameter recovery (0.06 to 0.074) as well as for the hit rate and Brier and SpS as measures of predictive accuracy (0.069 to 0.087). The amount of heterogeneity (factor 5) shows medium-to-large effect sizes on all measures of prediction accuracy (0.120 to 0.134) except RMSE(V), and a medium effect size on parameter recovery measured by RMSE (0.065). The model complexity that relates to the number of attributes and attribute levels except the brand constants (factor 3) also reveals medium effect sizes on all measures of parameter recovery (0.057 to 0.099), while the number of brands per nest shows medium effect sizes on all goodness-of-fit measures except PC (0.066 to 0.097) and medium-to-large effect sizes on all measures of prediction

Table 2 Effect sizes[a] of main and interaction effects on parameter recovery, goodness-of-fit and predictive accuracy measured by Eta squared (η^2)

Source (degrees of freedom)	Parameter recovery			Goodness-of-fit				Predictive accuracy				
	Correlation	RMSE	MAE	LL	PC	BS	SpS	Hit rate	BS	SpS	LL	RMSE(V)
Model (1)	0.009	**0.066**	**0.075**	0.020	0.040	0.015	0.016	0.000	0.000	0.000	0.005	0.044
Nest size (4)	0.003	0.011	0.010	0.015	0.014	0.013	0.015	0.031	0.025	0.027	0.020	0.011
Brands (1)	0.000	0.014	0.016	**0.097**	0.003	**0.066**	**0.069**	**0.072**	**0.135**	**0.088**	***0.160***	0.006
Complexity (3)	**0.082**	0.057	**0.099**	0.011	0.039	0.014	0.017	0.020	0.020	0.021	0.032	0.005
Similarity (3)	0.014	**0.060**	**0.074**	***0.269***	***0.266***	***0.280***	***0.296***	**0.087**	**0.069**	**0.078**	0.043	0.042
Heterogeneity (1)	0.031	**0.065**	0.032	0.011	0.010	0.005	0.002	**0.126**	**0.120**	**0.134**	**0.131**	0.044
Model × Heterogeneity	0.001	0.005	0.004	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.004
Model × Similarity	0.009	**0.062**	**0.080**	0.014	0.011	0.013	0.013	0.000	0.000	0.000	0.002	0.041
Model × Complexity	0.001	0.013	0.023	0.000	0.007	0.000	0.000	0.000	0.000	0.000	0.001	0.004
Model × Brands	0.001	0.004	0.004	0.001	0.002	0.000	0.001	0.000	0.000	0.000	0.000	0.002
Model × Nest size	0.001	0.001	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
Heterogeneity × Similarity	0.010	0.006	0.008	0.010	0.010	0.010	0.010	0.018	0.015	0.017	0.013	0.004
Heterogeneity × Complexity	0.006	0.001	0.003	0.003	0.002	0.002	0.002	0.005	0.005	0.006	0.006	0.002
Heterogeneity × Brands	0.011	0.006	0.005	0.000	0.000	0.000	0.000	0.003	0.000	0.004	0.009	0.005
Heterogeneity × Nest size	**0.064**	0.002	0.003	0.014	0.014	0.015	0.015	0.017	0.017	0.018	0.016	0.003
Similarity × Complexity	0.011	0.013	0.019	0.007	0.010	0.005	0.007	0.002	0.002	0.003	0.004	0.014
Similarity × Brands	0.002	0.001	0.002	0.009	0.003	0.001	0.009	0.001	0.005	0.001	0.002	0.000
Similarity × Nest size	0.011	0.031	0.031	0.052	0.052	**0.064**	**0.066**	0.028	0.021	0.024	0.016	0.025
Complexity × Brands	0.002	0.005	0.004	0.002	0.018	0.002	0.002	0.003	0.003	0.003	0.001	0.008
Complexity × Nest size	0.017	0.012	0.008	0.023	0.024	0.022	0.021	0.033	0.026	0.032	0.030	0.016
Brands × Nest size	0.001	0.003	0.001	0.005	0.003	0.002	0.005	0.005	0.002	0.005	0.005	0.002

[a]Bold and bold/italic values indicate effects of medium and large size, respectively

accuracy except RMSE(V) (0.072 to 0.160). Further, we observe medium effect sizes of the type of model on RMSE and MAE as measures of parameter recovery (0.066 to 0.075). The effect sizes for the number of respondents with a general preference towards higher quality versus lower quality brands of the two nests are consistently low (0.003 to 0.031), indicating that the partitioning of respondents across the two nests seems to have a negligible impact on the statistical model performance. Overall, the degree of similarity between alternatives in the two nests seems to be the most important factor, as it substantially affects all three types of performance measures (recovery, fit, and prediction), with medium to very large effect sizes for 9 out of 12 performance measures. Furthermore, sample heterogeneity and the number of brands per nest also seem to be primary drivers for the model performance, especially with regard to the predictive model performance. In contrast, the type of model and the model complexity only affect parameter recovery but do not show any noticeable effect sizes on model fit and predictive validity.

We further observe that most interaction effects have effect sizes close to zero, the rest (only five interactions) show medium-sized effects: the two interactions between the type of model and the similarity between alternatives in the nests on the RMSE (0.062) and MAE (0.080) as measures of parameter recovery, the interaction between sample heterogeneity and the number of respondents on the parameter recovery measured by the mean correlation (0.064), and the two interaction effects between the similarity of alternatives in the nests and the number of respondents on Brier and SpS as measures for goodness-of-fit (0.064/0.066).

One of the main findings we can derive from the results reported in Tables 2 and 6 and is that there seem to be no major differences between both types of models across treatments with regard to goodness-of-fit measures and in particular their ability to predict respondents' choice behavior, despite the underlying bimodal distribution of preference structures and the varying scenarios with respect to the correlations assumed for the unobserved portions of utilities between alternatives. It can therefore be concluded that the HB-MNL model is also able to uncover bimodal preference structures and to handle different similarities between alternatives within nests similarly well compared to the HB-NMNL model. The second major finding is that the similarity between alternatives in the nests (factor 4) plays a key role for the performance of a choice model both with regard to parameter recovery, model fit, and predictive validity. Furthermore, we observe (a) medium-sized main effects of the type of model and (b) medium-sized interaction effects between the type of model and the similarity between alternatives in the nests on parameter recovery measured by RMSE and MAE, which clearly suggests a different performance of the two models depending on the degree of similarity between alternatives in the nests. We comment on this finding below in more detail.

Table 3 reports the means of the twelve performance measures depending on the experimental condition (i.e., for the respective factor level). For each factor and performance measure the factor level means were tested for significant differences, and post hoc tests using the Bonferroni correction were conducted for the factors with more than two levels in order to analyze which of the factor level means do differ significantly from each other. For interpreting factor level differences, we again concentrate on those factors which showed at least a medium effect size ($\eta^2 \geq 0.06$).

Table 3 Means of performance measures by experimental condition (i.e., at the individual factor level)[a]

Factor	Parameter recovery			Goodness-of-fit				Predictive accuracy				
	Correlation	RMSE	MAE	LL	PC	BS	SpS	Hit rate	BS	SpS	LL	RMSE(V)
Model												
(1) MNL	0.950*	1.115*	0.732*	−0.482*	0.697*	−0.053*	0.852*	0.691	−0.086	0.750	−0.815*	2.481*
(2) NMNL	0.956*	0.826*	0.547*	−0.430*	0.742*	−0.048*	0.867*	0.695	−0.087	0.750	−0.848*	1.939*
Nest size												
(1) 125 875	0.953	1.044	0.676	−0.464	0.715	−0.051	0.857	0.677^{4,5*}	−0.090^{5*}	0.740^{5*}	−0.844^{5*}	2.384^{2*}
(2) 250 750	0.955	0.907	0.607	−0.493^{4,5*}	0.697^{4,5*}	−0.054^{5*}	0.847^{4,5*}	0.670^{4,5*}	−0.093^{4,5*}	0.732^{4,5*}	−0.879^{4,5*}	2.006^{1*}
(3) 500 500	0.954	0.930	0.617	−0.455	0.720	−0.050	0.860	0.692^{5*}	−0.086	0.748^{5*}	−0.840^{5*}	2.166
(4) 750 250	0.954	0.930	0.613	−0.444^{2*}	0.725^{2*}	−0.049	0.863^{2*}	0.707^{1,2*}	−0.084^{2*}	0.760^{2*}	−0.813^{2*}	2.168
(5) 875 125	0.949	1.042	0.685	−0.424^{2*}	0.739^{2*}	−0.047^{2*}	0.869^{2*}	0.721^{1,2,3*}	−0.080^{1,2*}	0.770^{1,2,3*}	−0.780^{1,2,3*}	2.325
Brands												
(1) 2	0.954	0.904*	0.598*	−0.398*	0.726*	−0.056*	0.874*	0.722*	−0.097*	0.775*	−0.738*	2.106*
(2) 3	0.953	1.037*	0.682*	−0.514*	0.713*	−0.045*	0.844*	0.664*	−0.077*	0.725*	−0.925*	2.313*
Individual parameters												
(1) A5L3	0.964^{3,4*}	1.187^{2,3,4*}	0.807^{2,3,4*}	−0.481^{4*}	0.698^{4*}	−0.053^{4*}	0.849^{4*}	0.712^{3,4*}	−0.081^{3,4*}	0.765^{3,4*}	−0.778^{3,4*}	2.221
(2) A5L5	0.961^{3,4*}	0.948^{1,4*}	0.645^{1,4*}	−0.448	0.717^{4*}	−0.050	0.861	0.703^{4*}	−0.084^{4*}	0.758^{4*}	−0.805^{3,4*}	2.070
(3) A9L3	0.949^{1,2,4*}	0.930^{1*}	0.585^{1*}	−0.466	0.706^{4*}	−0.051^{4*}	0.857^{4*}	0.683^{1*}	−0.090^{1*}	0.741^{1*}	−0.861^{1,2*}	2.227
(4) A9L5	0.939^{1,2,3*}	0.817^{1,2*}	0.522^{1,2*}	−0.429^{1*}	0.756^{1,2,3*}	−0.047^{1,3*}	0.870^{1,3*}	0.675^{1,2*}	−0.091^{1,2*}	0.735^{1,2*}	−0.881^{1,2*}	2.322
Similarity												
(1) 0202	0.955^{2*}	1.166^{2,3,4*}	0.766^{2,3,4*}	−0.320^{2,3,4*}	0.802^{2,3,4*}	−0.035^{2,3,4*}	0.904^{2,3,4*}	0.738^{2,3,4*}	−0.076^{2,3,4*}	0.783^{2,3,4*}	−0.765^{2,3,4*}	2.590^{2,3,4*}
(2) 0208	0.946^{1,3,4*}	0.908^{1,4*}	0.603^{1,3,4*}	−0.461^{1,4*}	0.716^{1,4*}	−0.052^{1,4*}	0.857^{1,4*}	0.689^{1,4*}	−0.087^{1,4*}	0.749^{1,4*}	−0.828^{1,4*}	2.050^{1*}
(3) 0802	0.955^{2*}	1.017^{1,4*}	0.673^{1,2,4*}	−0.451^{1,4*}	0.722^{1,4*}	−0.050^{1,4*}	0.862^{1,4*}	0.697^{1,4*}	−0.086^{1,4*}	0.752^{1,4*}	−0.830^{1,4*}	2.306^{1,4*}
(4) 0808	0.956^{2*}	0.791^{1,2,3*}	0.517^{1,2,3*}	−0.592^{1,2,3*}	0.637^{1,2,3*}	−0.065^{1,2,3*}	0.814^{1,2,3*}	0.649^{1,2,3*}	−0.097^{1,2,3*}	0.717^{1,2,3*}	−0.902^{1,2,3*}	1.893^{1,3*}
Heterogeneity												
(1) 005	0.959*	1.114*	0.700*	−0.437*	0.731*	−0.049*	0.862	0.731*	−0.077*	0.780*	−0.746*	2.481*
(2) 025	0.947*	0.827*	0.580*	−0.475*	0.708*	−0.052*	0.856	0.655*	−0.096*	0.719*	−0.916*	1.939*

[a]Superscripts on means refer to the factor levels and indicate significant differences at the 0.05 level (considering Bonferroni correction). Underlined values denote the factor levels showing the best performance w.r.t. the measure considered, respectively

As expected, less sample heterogeneity positively affects the predictive accuracy of the models (except for the RMSE(V)). Conversely, a higher amount of heterogeneity seems to decrease (i.e., improve) absolute error measures, both with respect to parameter recovery (RMSE, MAE) and prediction accuracy (RMSE(V)). The latter finding, however, might be an artefact concerning the way the part-worth utilities were generated. In particular, a higher amount of heterogeneity leads to a larger variance of the part-worth utilities on an individual respondent level. This in turn increases the impact of the deterministic part of the utility, and makes the choice behavior of respondents less stochastic.[10] Increasing the sample heterogeneity therefore has a similar effect as decreasing the error variance. As a consequence, MAE and RMSE values which penalize absolute deviations between "true" and re-estimated part-worths tend to get smaller (better) with a higher amount of heterogeneity. In contrast, the mean Pearson correlation measures the congruence between "true" and re-estimated part-worth utilities and therefore measures the goodness of parameter recovery in relative terms.

The effects of the log-sum parameters on goodness-of-fit measures are most noticeable (following the discussion above). In case of small log-sum parameters (high correlations between alternatives) in both nests, the model fit is significantly better as compared to the other settings. However, even if only one of the two log-sum parameters is small ($\lambda_1 = 0.2, \lambda_2 = 0.8$ or $\lambda_1 = 0.8, \lambda_2 = 0.2$) we observe a much better fit as compared to the setting that is most similar to a MNL scenario (small correlations between the stochastic components in both nests, i.e., $\lambda_1 = 0.8, \lambda_2 = 0.8$). Considering the interaction effects between the similarity of alternatives and the type of model it is noticeable that there are no mentionable effects (compare Table 2). Therefore, it can be concluded that a higher correlation between alternatives in at least one nest has a positive effect on the model fit independent of the type of model. The higher the correlation between alternatives within a nest, the more asymmetric are the substitution rates among alternatives in different nests. Thus, both models (and in particular the HB-MNL model that assumes independence between alternatives) seem to be able to soften the IIA property equally well.

Furthermore, we found medium effect sizes for the interaction between the similarity of alternatives and the nest sizes on Brier and SpS as measures of model fit (see Table 2). Table 7 (see Appendix 3) illustrates that a larger number of respondents in the nest with the higher correlation between alternatives (i.e., here the IIA property is strongly softened) tends to provide a better model fit (with $N_1 = 875 \text{ and } \lambda_1 = 0.2$ or $N_2 = 875 \text{ and } \lambda_2 = 0.2$ providing the best model fit). Again, this does not depend on the type of model. Since we observe nearly the same pattern of results for Brier and SpS independent of whether the higher correlation between alternatives refers to the high quality nest ($\lambda_1 = 0.2, \lambda_2 = 0.8$) or the low-quality nest ($\lambda_1 = 0.8, \lambda_2 = 0.2$) the predefined nest structure (high quality brands in the first nest vs. low-quality brands in the second nest) does not seem to influence the performance of the models.

[10] Concerning the degree of separation between components in segment-specific models, Vriens et al. (1996) and Andrews et al. (2002a, b) observed a similar effect. Further, almost identical effects are reported in Wirth (2010).

In addition, small log-sum parameters lead to significantly better hit rates, BS, and SpS in the validation sample (compare Table 3).

Somewhat different implications result for the impact of the degree of similarity among alternatives on parameter recovery when measured in absolute terms (RMSE, MAE). Considering the two main effects, less correlation between alternatives in the nests enables a better parameter recovery (compare Table 3). Table 8 in the Appendix (Appendix 3: Additional Tables) further shows the related interactions between the degree of similarity of alternatives and the type of model. Here, the interaction effects clearly suggest that the HB-MNL model performs increasingly worse as correlations in at least one nest are higher ($\lambda_1 = 0.2$ and/or $\lambda_2 = 0.2$), while the HB-NMNL model is only marginally affected by the size of the log-sum parameters and adapts to the degree of similarity between alternatives, as expected. For $\lambda_1 = 0.8$ and $\lambda_2 = 0.8$ the HB-NMNL and the HB-MNL models perform equally well (note that when the log-sum parameters approximate 1, the NMNL model degenerates to a MNL model). Consequently, it can be assumed that the HB-NMNL model has advantages in terms of parameter recovery if nested structures with high correlations are present.

An increasing model complexity, i.e. a larger number of parameters (factor 3), slightly worsens parameter recovery when measured in relative terms (mean correlations) but at the same time improves parameter recovery when measured in absolute terms (RMSE, MAE). The latter result is not obvious at first glance and seems again to be an artefact concerning the way the part-worth utilities were generated. As mentioned in the section on data generation above, we leaned on the Monte Carlo study of Vriens et al. (1996) and generated "true" mean part-worth utilities of respondents according to a uniform distribution with a maximum utility range of -10.2 to $+10.2$ across attributes. However, the more attributes and/or attribute levels given that fixed utility range, the more parameters to be estimated and the smaller the corresponding intervals for the attributes (excluding for brands and prices, compare section on data generation). As a consequence, the generated "true" mean part-worth utilities tend to be smaller in their magnitudes which in turn leads to smaller absolute deviations between "true" and re-estimated mean part-worth utilities as reported by the RMSE and MAE measures for parameter recovery. In contrast, the mean correlation measures the congruence between "true" and re-estimated part-worth utilities and is not affected by a decrease in absolute deviations.

Further, an improvement in fit and especially in prediction accuracy can be observed in case of two instead of three brands per nest and thus smaller choice sets (except for the BS). This tendency was expected because it is easier to predict more accurately for fewer alternatives per choice task. The probability of predicting first choices correctly by the null model decreases with more alternatives per choice task. Note, however, that the relative hit rates are 2.88 (two brands per nest) and 3.98 (three brands per nest) times higher as compared to the null model.[11] A similar reasoning applies to the BS. The numerator compares chosen alternatives ($Y_{njc} = 1$) and non-chosen alternatives ($Y_{njc} = 0$) to the estimated probabilities (cf. Eq. (9)). Hence, the

[11] This is because the probability of predicting first choices correctly by the null model decreases with more alternatives per choice task. The hit rate of the null model with four alternatives (two alternatives per nest) is 0.25, and the hit rate of the null model with six alternatives (three alternatives per nest) is 0.17.

BS exploits more information of the predictive distribution of choice probabilities than both the LL and the SpS. Note that the part of the sum that takes into account non-chosen alternatives $(Y_{njc} = 0)$ decreases in the null model for an increasing number of alternatives per choice set. In the null model, this relative decrease in case of larger choice sets is greater than the relative increase in the part of the sum that takes into account chosen alternatives $(Y_{njc} = 1)$ in case of smaller choice sets.

Up to this point, the most important finding is that the HB-MNL model fits and predicts equally well for the considered nested structures compared to the HB-NMNL model. All effect sizes of the type of model on the related fit and predictive validity measures are small if not negligible. On the other hand, the HB-NMNL model shows the expected advantages as far as parameter recovery is concerned and in particular provides lower absolute parameter recovery errors.[12]

In Table 4 we summarize the main results about effect sizes of the type of model and the different experimental factors as well as about how the model performance is generally influenced by a manipulation of factor levels.

Table 4 Overview of main results

Factor	Effect size	General tendencies
Model	Medium: parameter recovery (RMSE, MAE)	The HB-NMNL model improves parameter recovery over the HB-MNL model (RMSE, MAE)
Nest size (number of respondents with a general tendency in their preference towards a nest)	Small or near zero: for all performance measures	The partitioning of respondents across the nests does not affect the performance of both models
Number of alternatives per nest (brands)	Large: predictive accuracy (LL) Medium: goodness-of-fit (LL, BS, SpS), predictive accuracy (hit rate, BS, SpS)	Increasing the number of alternatives per nest • decreases predictive accuracy (hit rate, SpS, LL) • decreases goodness-of-fit (LL, SpS) • improves BS (goodness-of-fit, predictive accuracy)
Model complexity (number of parameters excluding brand constants)	Medium: all parameter recovery measures	Increasing the number of parameters • improves absolute parameter recovery (RMSE, MAE) • decreases relative parameter recovery (correlation)
Similarity between alternatives (log-sum coefficients)	Very large: all goodness-of-fit measures Medium: parameter recovery (RMSE, MAE), predictive accuracy (hit rate, BS, SpS)	A high similarity in both nests • improves goodness-of-fit • improves predictive accuracy (hit rate, BS, SpS) A high similarity in at least one nest decreases parameter recovery (RMSE, MAE) of the HB-MNL model
Preference heterogeneity	Medium: parameter recovery (RMSE), predictive accuracy (hit rate, BS, SpS, LL)	More heterogeneity • improves parameter recovery (RMSE) • decreases predictive accuracy (hit rate, BS, SpS, LL)

[12] For the sake of completeness, Table 9 in the Appendix (Appendix 3: Additional Tables) summarizes the interaction effect between the nest size and the sample heterogeneity on the mean Pearson correlation (parameter recovery).

R and SAS scripts for data generation, estimation, and validation procedures are available as supplementary materials via the Open Science Framework (OSF), see at https://osf.io/9afxc/.

4 Empirical Study

Without loss of generality, we considered nested market structures in the form of two submarkets in our Monte Carlo study, representing a high-price/high-quality and a low-price/low-quality tier. Many real-world markets show this two submarket structure, as will be detailed in the following. The specific structure, i.e. whether two price-quality tiers or more granular structures including ‘low’, ‘mid’, and ‘high’ tiers exist, depends heavily on the product category. In the automotive market, for instance, nearly every vehicle class contains both low-tier and high-tier options. Manufacturers oftentimes differentiate offerings within a class across two, three, or sometimes more tiers, depending on the car brand. Similar patterns are observed in other markets, such as laundry detergents, where high-quality, high-price products coexist with lower-quality, more affordable ones − and depending on the brand sometimes alongside an additional mid-tier. In the telecommunications market, premium plans (e.g., €90/month 5G unlimited) exist alongside basic, low-cost options (e.g., €10/month limited data and call minutes), constituting a two-submarket structure. In the car tire market, as considered in our empirical study, higher-priced premium tire brands are offered along with lower priced non-premium brands, again speaking for a two-nest structure.

Integrating these tiered structures into a model allows for a more flexible and behaviorally realistic assignment of respondents to preference segments − or “nests” − *after* their choices and preference patterns are observed. Pre-assigning respondent preferences in practical applications to these nests before model estimation would require substantial prior knowledge and the assumption that respondents’ preferences are fixed and known in advance. This assumption overlooks the possibility that respondents might update or reorient their preferences when exposed to the alternatives presented in the choice tasks. In contrast, *nested logit models* allow for a probabilistic assignment of respondents’ preferences to nests during model estimation, which captures this uncertainty more appropriately. In a Bayesian framework, this flexibility is even more pronounced: respondents can switch between nests across iterations, and only at convergence are they assigned to a specific segment. This reflects the dynamic nature of preference learning, as formally indicated by the conceptual separation of the error term of the nested logit model into a cross-sectional heterogeneity part and a dynamic (learning) part, compare Eq. (5) in Sect. 2.1. Actually, nested market structures and the related preference heterogeneity are often only revealed during the survey process itself. It is through respondents’ observed choices that the relevance of tiers becomes empirically evident. Therefore, an a priori segmentation of respondents according to their preferences for different nests followed by a separate nest-specific modeling would be premature. Only once these structures are validated through initial data analysis could such an a priori segmentation of respondents be justified. Even in this case can a single nested model capture this complexity more efficiently within a unified framework compared to splitting the data and estimating separate models, at least from a theoretical perspective.

In our Monte Carlo study, we found that parameter recovery could be substantially improved by the more flexible HB-NMNL model compared to the HB-MNL model, as opposed to model fit and predictive accuracy (compare Sect. 3). Therefore, more flexibility need not necessarily pay off in terms of a better model performance in all dimensions. At this point, it is important to note that parameter recovery cannot be evaluated in empirical studies, since true preferences of respondents are unknown there. Hence, for parameter recovery which is essential for product design decisions, Monte Carlo studies are the more important. In the following, we will therefore compare the HB-NMNL and HB-MNL based on an empirical data set for model fit and predictive validity, and further to uncover whether this market has a nested structure.

The summer tires data for our empirical study was made available by Peter Kurz, formerly Head of Research and Development at Kantar TNS, Munich, Germany. The data covers stated choices for 4,026 respondents in the year 2016, collected in France, Germany, Italy, Spain and the United Kingdom via a discrete choice experiment. At that time, the summer tire market was made up of offerings by 10 competing summer tire brands (in alphabetical order: Bridgestone, client brand, Continental, Firestone, Goodyear, Hankook, Kleber, Michelin, Pirelli, and a further low price brand). The client brand was the customer of the market research institute for this study, and to maintain the anonymity of the client brand a second brand had to be anynomized. In this labelled choice experiment, respondents were asked to choose their preferred brand from 15 choice sets, where each choice set included three alternatives plus a no-purchase option and each alternative was described along 8 attributes with 3 to 10 attribute levels: *brand* (10 levels), *tire type* (standard, comfort, sports), *longevity* (−8,000 km, −4,000 km, standard, +4,000 km, +8,000 km), *rolling resistance/fuel consumption* and *grip on wet roads* (A, B, C, E, F),[13] *consumer reviews* and *independent test results* (not available, 1 to 5 stars),[14] and price (7 levels). To prevent that premium (non-premium) brands were displayed with unrealistic low (high) price levels, i.e. to adequately represent different tire qualities in price, an alternative-specific design with brand-specific price attributes was generated for this experiment. For example, the price levels for the most expensive brand Michelin ranged between 77€ to 101€, while for the least expensive brand, the low-price brand (anonymized), prices were between 49€ to 73€. More details on the data set can be found in Goeken et al. (2021), where the data was used in a different context, namely to illustrate the estimation of different hierarchical Bayesian conjoint choice models, such as latent class, mixture-of-normals and Dirichlet process prior models.

In a first approach, it seemed reasonable to assume that the market could best be characterized to consist of two submarkets, one nest with higher-priced premium brands (Bridgestone, Continental, Goodyear, Michelin, and Pirelli) and a second nest with lower-priced non-premium brands (client brand, Firestone, Hankook, Kleber, and the anonymized low price brand). Based on a total of 60,390 choices, we estimated the HB-MNL and HB-NMNL models and compared them for model fit and

[13] The levels for these two attributes represent the EU tire labels of the European Union.

[14] The levels for these two attributes represent the established 5-star rating system for product quality.

predictive performance. Note that the no-purchase option was not assigned to one of the product nests but was treated as a separate nest in the HB-NMNL.[15] To assess predictive validity, we randomly selected one of the 15 choice sets of each respondent to constitute a validation sample, and we replicated this procedure a second time to avoid a systematic error due to a one-time selection bias.[16] Model fit was evaluated by the log-likelihood (LL), the percent certainty (PC) and the in-sample hit rate (IHR), based on the entire sample. Predictive performance was assessed by the out-of-sample hit rate (OHR), averaged over the two randomly generated validation samples. Table 5 summarizes the results for model fit and predictive accuracy in terms of 95% credible intervals for the four measures.

The percent certainty as well as the in- and out-of sample hit rates suggest a good fit and predictive performance of both models. A percent certainty between 0.2 and 0.4 already indicates a good model fit. Note that the out-of-sample hit rate is lower than the in-sample hit rate as expected, but still more than two times higher than the chance rate of 25%, given choice sets with 4 alternatives (including the no-purchase alternative). Importantly, the 95% credible intervals for each of the performance measures strongly overlap, implying no statistical difference in model fit and predictive accuracy between the two models. A closer look on the parameter estimates reveals that the log-sum coefficients for both nests turn out slightly higher than 1, with 95% credible intervals of [1.074;1.158] and [1.001;1.098] for λ_1 (premium brands) and λ_2 (non-premium brands), respectively. Given no convergence problems of the HB-NMNL model, this suggests that the predefined nest structure does not seem to represent the true hierarchical decision process of the respondents. In other words, the groups of alternatives inside the nests might be not homogeneous enough. It is important to note that the estimated log-sum coefficients are so close to 1 that the exceedance of the theoretical upper limit of 1 can be attributed to estimation error, i.e. there is actually no correlation among the stochastic components such that the NMNL model more or less degenerates to the MNL model. For this reason, the two models perform equally well in fit and predictive performance, which can additionally be seen as indicative for a proper estimation of the HB-NMNL model.[17]

We therefore tested another two conceivable nested structures. Based on a finer differentiation of the observed price ranges of the brands, we defined a 3-nest-composition representing three price tiers, consisting of one high-price nest (*first nest:* Continental, Michelin; 69€ to 101€), one mid-price nest (*second nest:* Bridgestone,

Table 5 Goodness-of-fit and predictive accuracy measures by model type. Represented are the 95% credible intervals of the posterior distributions

Model	LL	PC	IHR	OHR
HB-MNL	[−42,776.42;−41,953.70]	[0.489;0.499]	[0.705;0.711]	[0.526;0.545]
HB-NMNL	[−42,892.14;−42,313.14]	[0.488;0.495]	[0.702;0.710]	[0.527;0.543]

[15] This also holds true for the different 3-nest-structures tested subsequently (see below).

[16] Note that no fixed holdout choice tasks were collected for this study. Given the large sample size of more than 4,000 respondents, predictive validity could also have been assessed by splitting up the respondents into two halves and using 50% of them for estimation and 50% of them for validation.

[17] Note that without exception all re-estimated log-sum coefficients in our Monte Carlo study fell into the interval [0,1], which was expected given the explicit generation of nested market structures there.

Client Brand, Goodyear, Kleber, and Pirelli; 57€ to 89€), and one low-price nest (*third nest:* Firestone, Hankook, low price brand; 49€ to 77€). Second, we specified a 3-nest structure relating to the *tire type* attribute (*comfort:* Continental, Goodyear, Michelin; *sport:* Bridgestone, Pirelli; *standard:* Client Brand, Firestone, Hankook, Kleber, low price brand). Again, the "true" substitution pattern (if one existed) could not be approximated better; all log-sum coefficients turned out again close to 1 for both 3-nest structures. As a result, for the summer tires product category, consumers did either not follow a hierarchical decision process such that all tire brands were perceived as equal substitutes or they did follow one which, however, was not obvious. At this point, please remember that the less complex HB-MNL model performed comparably well in the Monte Carlo study regarding model fit and in particular predictive accuracy, given a wide variety of substitution patterns in 2-nest markets, i.e. even in markets where a clear nest structure existed. Future studies should therefore investigate whether the property of the HB-MNL model to uncover even nested market structures (if these existed) generalizes to empirical data, too.

Overall, as a preliminary conclusion from our studies and at least if new product (line) design is of primary concern, we recommend the estimation of the HB-NMNL since our Monte Carlo study revealed a better parameter recovery for this model compared to the simple HB-MNL. Whether a nested market structure exists then becomes evident from the estimates for the log-sum coefficients. As a second important conclusion from our empirical study, questions regarding a potential hierarchical decision process of consumers should be included in a discrete choice experiment in order to reveal a possible nested market structure, if a priori knowledge about it does not exist. Of course, due to the current lack of commercial software for the HB-NMNL model, basic expertise in *R* or *Stan* programming is required.

5 Summary, discussion, and conclusions

Previous Monte Carlo studies related to traditional conjoint data or stated choice data mainly focused on the comparison of the statistical performance of different competing models. In particular, many of these models have been implemented in a hierarchical Bayesian estimation framework because researchers found out that analyzing data at an aggregate level could obscure important aspects. In order to address heterogeneity in discrete choice models, the MNL model estimated with HB, also known as HB-MNL model, became the state-of-the-art in marketing applications. Nevertheless, the MNL model is subject to the IIA property. Although the NMNL model is often suggested as a remedy for relaxing the IIA property, to the best of our knowledge no Monte Carlo study related to stated choice data has yet systematically explored the performance of HB-NMNL models. The goal of the present study was to compare the performance of HB-MNL models and HB-NMNL models under different nested structures. In particular, we wanted to investigate how robust the HB-MNL model works against violations of the underlying assumptions (here: nested structures in the data). Therefore, the statistical performance of the HB-MNL regarding parameter recovery, model fit and predictive validity was of special interest (as compared to the HB-NMNL model). To evaluate the comparative performance (on the dimensions parameter recovery, goodness-of-fit and predictive accuracy) of the HB-MNL versus the HB-NMNL we experimentally manipulated five factors: (1) the number/partitioning of respondents with a general preference towards either nest, (2) the number of alternatives (brands) offered per nest, (3) the number of parameters to be estimated at the individual respondent level (in addition to the brand constants for the alternatives), (4) the degree of similarity between alternatives, and (5) the level of preference heterogeneity. Main effects and first-order interaction effects were analyzed using analysis of variance.

One of the main findings we could derive from our Monte Carlo study is that there seem to be no major differences between both types of models with regard to goodness-of-fit measures and in particular their ability to predict respondents' choice behavior, despite the underlying bimodal distribution of preference structures and the varying scenarios with respect to the correlations assumed for the unobserved portions of utilities between alternatives. It can therefore be concluded that the HB-MNL model is also able to uncover bimodal preference structures and to handle different similarities between alternatives within nests similarly well compared to the HB-NMNL model. The second major finding is that the similarity between alternatives in the nests plays a key role for the performance of the choice models with regard to all performance dimensions, namely parameter recovery, model fit and predictive validity. Furthermore, we observed (a) noticeable effects of the type of model (HB-MNL vs. HB-NMNL) and (b) noticeable interaction effects between the type of model and the similarity between alternatives in the nests on parameter recovery measured in absolute terms (RMSE, MAE). This finding clearly suggests a different performance of the two models depending on the degree of similarity between alternatives in the nests. While the HB-NMNL model is only marginally affected by different log-sum coefficients, we found out that the HB-MNL model performs increasingly worse when correlation in at least one nest is higher. As expected, the HB-NMNL model adapts to the degree of similarity between alternatives. For small similarity between alternatives in both nests the HB-NMNL and the HB-MNL models perform equally well (note that when the log-sum parameters approximate 1, the NMNL model degenerates to a MNL model). Consequently, it can be assumed that the HB-NMNL model shows advantages as far as parameter recovery is concerned. Further, sample heterogeneity and the number of brands per nest also seem to be primary drivers for the model performance, especially with regard to the predictive model performance. In contrast, the type of model and the model complexity only affect parameter recovery but do not show any noticeable effects on model fit and predictive validity.

Summing up, our results from the Monte Carlo study show that the choice of the model only has a moderate influence on our measures of performance. Although both models are supplemented with a unimodal prior distribution, we can assume that the HB-MNL model is also able to uncover bimodal preference structures and to handle different similarities between alternatives within nests similarly well compared to the HB-NMNL model. In the presence of a nested structure with high correlations between alternatives the HB-NMNL model shows advantages compared to the HB-MNL model in terms of parameter recovery measured in absolute terms.

We further compared the two models in an empirical study with stated choice data for summer tires. On the one hand, this study revealed that it seems recommendable to apply the HB-NMNL model in a first step to check if any reasonable nest structure (if a priori not yet validated) actually exists. If one existed and product design decisions are the primary focus of a study, the use of the HB-NMNL model might be advantageous given the evidence of its superior parameter recovery in nested markets found in our Monte Carlo study. If prediction is the objective, the use of the less complex HB-MNL might be sufficient, provided its comparably good predictive performance is generalizable to empirical data, too. As mentioned before, we expect this in the light of our experiences with the various types of nested market structures investigated in our simulation study. Yet, more comparisons of the two models based on empirical data should be conducted in the future to check the robustness of our findings from the Monte Carlo study regarding predictive performance.

Finally we want to note a few more limitations. A Monte Carlo study does not necessarily reflect the real behavior of respondents. A certain number of parameters were varied, whereby some parameters cannot be varied in practical choice modelling applications. In particular, the part-worth utility structure, the nest sizes, the similarity between alternatives (the log-sum coefficients) and the preference heterogeneity cannot be modified by the researcher. Some factors (e.g., the log-sum coefficients, the nest sizes, or the part-worth utility structure) can only be influenced by the determination of an adequate substitution pattern that reflects the choice behavior of respondents. In this study, the case was considered that part-worth utilities follow a normal distribution in both nests. In future research, the influence of different preference distributions could be investigated (e.g., Andrews et al. 2002b; Otter et al. 2004). The amount of preference heterogeneity was simulated by a covariance matrix with equal diagonal elements (0.05 for a homogenous sample structure and 0.25 for a heterogeneous sample structure). Empirical studies show that with regard to some attributes respondents reveal rather homogeneous preferences and perceive other attributes as very different. Based on this, Wirth (2010) adapted the distribution of the diagonal elements in his Monte Carlo study. He generated a distribution of variances from which he randomly drew the corresponding diagonal elements. However, using that kind of heterogeneity would have masked our supposed bimodal part-worth structure. In future research it might be interesting to adjust the distribution of variances for part-worth utilities or the supposed utility structure to account for this kind of more realistic preference heterogeneity. In addition, the log-sum coefficients were assumed to be fixed parameters over all respondents. Empirical studies show that respondents might differ in the perception of similarity between alternatives. Bhat (1997) allowed for varying log-sum parameters across respondents by defining a continuous monotonically increasing function that maps to the interval [0,1]. The function $F(\alpha_0 + \alpha' c_n) = \lambda_n$ transforms socio-demographic characteristics c_n (e.g., income) of respondent n to individual log-sum coefficients λ_n that must lie in the 0–1 range. Empirical results show that accounting for log-sum heterogeneity leads to a better fit and better parameter recovery. Further, in future research it might be worth to consider different prior distributions. The unimodal prior distribution tends to shrink units towards the population mean, which could mask multimodal preference structures. A mixture of several multivariate normal distributions (Allenby et al. 1998) offers additional flexibility and could improve the estimates. Thus, the question arises how the comparative performance of the HB-MNL model versus the HB-NMNL will be affected when the prior distributions account for multimodality. Although in this study an interaction between prices and brands was already considered by the two nests, it might also be interesting to include additional interaction terms between attributes in both models (e.g., between price and brand). Not least, we limited our Monte Carlo study to the existence of two nests. Future research could extend our study by adding a third or even a fourth nest (and/or including more items in the nests) to represent other realistic nested market structures in order to check for further generalizability of our findings.

The NMNL model relaxes the IIA property by grouping nearly substitutable alternatives. Nevertheless, the ratio of choice probabilities of two alternatives belonging to different nests only depends on alternatives in the same nests. In the literature, this effect is referred to as IIN property. Another approach to relax the IIA property and additionally the IIN property is the MNP model. The MNP model allows a completely free correlation structure among the unobserved part of utilities, but has no closed-form mathematical structure. This leads to a considerable more demanding estimation process. Further limitations of the NMNL model should be considered. Sivakumar (1995) found out that the probability of choosing a brand from a cluster (nest) increases with the cluster (nest) size. Especially, this cluster effect could be observed in high-priced, high-quality clusters. Although a Monte Carlo study is not affected by a cluster effect, the number of alternatives offered per nest were chosen to be equal in our study (2 or 3). However, Sivakumar (1995) proposed a possibility to adjust the inclusive value, the log of the sum of exponentiated representative utilities, to reduce a possible cluster effect. Another limitation of the NMNL model is the difficulty to determine an adequate substitution pattern ex ante that reflects the complex choice behavior of respondents. Examples for appearing problems when determining nested structures can be found in Train et al. (1989), Lee (1999), Elshiewy et al. (2017) and Train et al. (1987).

Considering these results, it can be concluded that, in practice, it is a matter of opinion and personal preference whether a market researcher prefers to use a HB-MNL or HB-NMNL model. As mentioned above, an “adequate” or “proper” substitution pattern needs to be determined for the HB-NMNL model prior to estimation. Consequently, the superior parameter recovery of the HB-NMNL model in the presence of highly correlated nested structures contrasts with the more complex estimation process which is not as straightforward as that for the HB-MNL model, at least as long the HB-NMNL model has not yet been implemented in commercial software packages for choice modelling studies based on stated choice data (e.g., Sawtooth Software). Until then, one idea could be to include indicator variables for an a priori specified nest structure (e.g. higher-priced high quality vs. lower-priced low quality brands) as additional covariates in the HB-MNL model, as proposed by Grigolon and Verboven (2014) for their heterogeneous MNL model. Given that the HB-NMNL model did not show a superior predictive performance over the HB-MNL model in our study, this simple yet much less complex approach seems promising. Future research should further investigate whether this extended HB-MNL model can also improve the parameter recovery of the HB-MNL model for nested market structures.

Finally, we did not find an appropriate nesting structure in our empirical study on summer tires, all tested nest variants did not improve over the simpler HB-MNL. Picking up our argument about order effects as a rationale for correlated errors across choice sets, another possible nesting structure might be based on a position effect with respect to the three alternatives offered to respondents in each choice set. The position effect is also referred to as left–right bias due to the fact that most cultures process information from left to right (e.g. Boxebeld 2024). Accordingly, alternatives one and two, the most left ones, might have been closer substitutes for (parts of) the respondents compared to the third alternative presented in the third column. On the other hand, alternatives two and three might have been perceived as closer substitutes, if respondents went carefully through the whole choice set and the last impressions were more influential for the respondents' choice behavior (following Boxebeld 2024, p. 2). We leave this issue for future research.

Appendices

Appendix 1: Model estimation

Remember that the utility of respondent n for alternative j in choice situation s can be written as $U_{njs} = x_{njs}\beta_n + \varepsilon_{njs}$, with ε_{njs} being jointly distributed as a GEV. Suppose that N respondents and S choice situations/choice sets are considered. The vector $y_n = (y_{n1}, \ldots, y_{nS})^T$ then provides information about the alternative respondent n has chosen from choice set $1, \ldots, S$, respectively. The matrix $Y = (y_1, \ldots, y_N)$ contains all choice decisions for respondents $1, \ldots, N$. The vector $\lambda_n = (\lambda_{n1}, \ldots, \lambda_{nS})^T$ contains the corresponding log-sum coefficients whereas $\lambda_{ns} \in \lambda = (\lambda_1, \ldots, \lambda_K)\ \forall n, s$ and accordingly $B_{ns} \in (B_1, \ldots, B_K)\ \forall n, s$ provides information about the corresponding nest. Therefore, the likelihood for respondent *n* is:

$$L\left(y_n \middle| \beta_n, \lambda\right) = \prod\nolimits_s \frac{e^{\beta_n' x_{ny_{ns}s}/\lambda_{ns}} \left(\sum_{j\epsilon B_{ns}} e^{\beta_n' x_{njs}/\lambda_{ns}}\right)^{\lambda_{ns}-1}}{\sum_{l=1}^{K} \left(\sum_{h\epsilon B_l} e^{\beta_n' x_{nhs}/\lambda_l}\right)^{\lambda_l}}. \tag{A1}$$

The mixed nested logit choice probability[18] is the probability conditional on the fixed coefficients λ and the population parameters/hyperparameters b and W:

$$L\left(y_n \middle| b, W, \lambda\right) = \int L\left(y_n \middle| \beta_n, \lambda\right) p(\beta_n \,|b, W)\, d\beta_n. \tag{A2}$$

$p(\beta_n \,|b, W)$ represents the prior distribution of β_n and here a multivariate normal distribution with mean b and covariance matrix W. For Bayesian analysis further

[18] It is a product of nested logit models mixed over a density of coefficients. For this reason, the probability is called mixed nested logit choice probability.

(hyper) prior distributions on the model parameters b, W are needed, denoted as $p\,(b)$ and $p\,(W)$, respectively:

$$\begin{array}{lll} \beta_n & \sim & N\,(b, W)\,, \\ b & \sim & N\,(b_0, B_0)\,, \\ W & \sim & IW\,(v_0, S_0) \end{array} \tag{A3}$$

v_0 denotes the degrees of freedom and S_0 is the scale matrix, a d-dimensional identity matrix times the prior variance s_0, where d is the dimension of the data. A diffuse prior with a sufficient large variance is assumed for b. We use Gibbs sampling for Bayesian inference, which results in the following conditional posteriors:

$$\begin{aligned} & p(\beta_n|\lambda, b, W, y_n) \propto \prod_s \frac{e^{\beta_n' x_{ny_{ns}s}/\lambda_{ns}} \left(\sum_{j\varepsilon B_{ns}} e^{\beta_n' x_{njs}/\lambda_{ns}}\right)^{\lambda_{ns}-1}}{\sum_{l=1}^{K} \left(\sum_{h\varepsilon B_l} e^{\beta_n' x_{nhs}/\lambda_l}\right)^{\lambda_l}} p(\beta_n|b, W), \\ & p\,(b|W, \beta_n) = N\left(\overline{\beta}, W/N\right), \\ & p\,(W|b, \beta_n) = IW\left(v_0 + N, \frac{v_0 S_0 + N\overline{S}}{v_0 + N}\right), \text{ where, } \overline{S} = \sum_n (\beta_n - b)\,(\beta_n - b)^T /N, \\ & p\,(\lambda|\beta_n) \propto \prod_n L(y_n|\beta_n, \lambda). \end{aligned} \tag{A4}$$

Following Train (2001), the addition of the fixed effects (here: the log-sum coefficients $\lambda_1, \ldots, \lambda_K$) requires a fourth layer in the Gibbs sampling. Drawing the fixed coefficients simultaneously with the individual part-worth utilities would cause the fixed effects to be random across respondents, which is inconsistent with the assumption of fixed effects. Therefore, the estimation of fixed effects needs to be separated from that of random effects. Accordingly, the log-sum coefficients are determined separately in the fourth layer of the model conditional on the current values of the random coefficients. Note that there is generally no strict need to specify an additional prior for fixed effects, which is also the default setting in the RSGHB package we used for model estimation. Therefore, the log-sum coefficients were estimated based only on the likelihood function in this fourth layer.

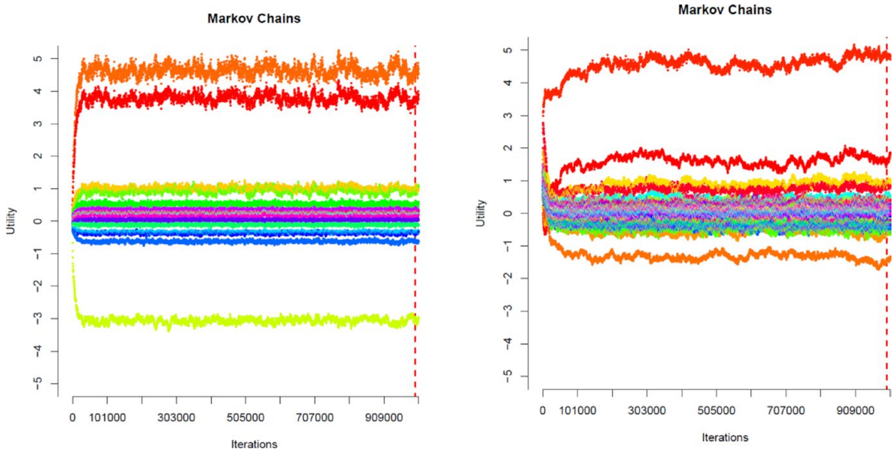
Draws from multivariate normal distributions and inverted Wishart distributions can be obtained conveniently as exemplary described in e.g. Train (2009). To draw from the conditional posterior on β_n, the Metropolis–Hastings algorithm is used. The resulting values β_n, λ, b and W converge to draws from the joint posterior distribution. We set the prior variance s_0 to 2 and the degrees of freedom v_0 to 5 plus the number of parameters. Hence, these settings are comparable to the default settings of Sawtooth Software (2016). To ensure convergence of the Markov chain the MCMC sampler was run for 1,010,000 iterations. In each case, we used a burn-in period of 1,000,000 iterations. In most cases, convergence could be determined much earlier. However, the very high number of iterations is used to ensure convergence in all cases. To reduce possible correlation among the draws, we only used every 10th draw of the remaining 10,000 draws to obtain point estimates. To control the estimation process, we monitored time-series plots and checked goodness-of-fit measures. Furthermore, we calculated Gelman and Rubin's potential scale reduction factor (Gelman and Rubin 1992).

Assuming a MNL model ($\lambda_k = 1 \ \forall k$) the likelihood as stated in Eq. (A5) simply degenerates to the product of standard logit formulas:

$$L\left(y_n \middle| \beta_n\right) = \prod\nolimits_s \frac{e^{\beta_n' x_{n y_{ns} s}}}{\sum_h e^{\beta_n' x_{nhs}}}. \tag{A5}$$

Estimation of the MNL model follows the same Gibbs sampling procedure but without the fixed coefficients λ, hence without the fourth layer. Note that estimation of a model with both fixed and random coefficients is expected to take about twice as much time than without the fourth layer for the fixed coefficients (cf. Train 2001, pp. 7 and Train 2009, pp. 320).

The following figure shows two selected trace plots for estimated HB-NMNL models, one for equal nest sizes ($N_1 = 500, N_2 = 500$), two alternatives per nest, 5 attributes with 3 levels each, high similarity between alternatives in the two nests ($\lambda_1 = \lambda_2 = 0.2$), and low preference heterogeneity (left-hand plot), and the other for unequal nest sizes ($N_1 = 875, N_2 = 125$), two alternatives per nest, 9 attributes with 3 levels each, high resp. low similarity between alternatives in nest 1 resp. nest 2 ($\lambda_1 = 0.2, \lambda_2 = 0.8$), and large preference heterogeneity (right-hand plot). As could be expected, convergence lasted longer for the model with the higher model complexity (9 attributes, right-hand plot). Given the true log-sum coefficients $\lambda_1 = 0.2 \text{ and } \lambda_2 = 0.2$ (left-hand plot) and $\lambda_1 = 0.2 \text{ and } \lambda_2 = 0.8$ (right-hand plot), we obtained $\widehat{\lambda}_1 = 0.19 \text{ and } \widehat{\lambda}_2 = 0.29$ and $\widehat{\lambda}_1 = 0.43 \text{ and } \widehat{\lambda}_2 = 0.88$ as model estimates, respectively. Further, the upper two and the lower one traces in both treatments (plots) refer to the estimated brand constants for the two premium brands and the highest price level.



In a recent study, van Horn (2024) compared the Metropolis Hastings (MH) and the Hamiltonian Monte Carlo (HMC) algorithms for estimating the HB-MNL model along six stated choice data sets, with the number of respondents, choice tasks, and estimated parameters ranging between 600 and 6,000, 4 and 22, and 15 to 27. In a nutshell, as summarized by Orme (2024), MH and HMC provided similarly good holdout predictions and well-matched posterior estimates of means and standard deviations given a comparable amount of computing time, although HMC revealed superior convergence diagnostics (Rhat, effective sample size). Since early HMC iterations take much longer than later HMC iterations, MH might be preferable for small and moderate-sized discrete choice data sets, while HMC for complex data sets with a very large number of estimated parameters (> 100), see Orme (2024), p. 4.

Appendix 2: Details on the data generation process

Design: Since all five or nine non-brand attributes had either three or five levels and each brand appeared in each choice set so that the number of alternatives per choice set was 4 or 6, the least complex model comprised 13 individual-level parameters (4 brands and 5 non-brand attributes à 3 levels), while the most complex model 41 individual-level parameters (6 brands and 9 non-brand attributes à 5 levels), compare Table 1. As further mentioned in Sect. 2.2, each two out of the four alternatives or each three out of the six alternatives (depending on the setting of factor 2) were further defined as high-quality brands respectively low-quality brands, and one of the five or nine attributes (depending on the setting of factor 3) was chosen to represent the price attribute with three or five price levels. This price attribute was constrained to be an ordinal variable where the first attribute level referred to the lowest price level and the last attribute level to the highest price level. In order to obtain a nested structure, the final design was generated in such a way that higher prices tended to occur together with high-quality brands, and lower prices were assigned to low-quality products. For these specific price-quality patterns, occurrence probabilities had to be defined. For the treatments with three attribute levels (see Table 1, factor 3, A5L3 and A9L3), explorative searches during the data generation process provided evidence that an incidence rate of 90% for the highest (lowest) price level and 10% for the medium price level for the high-quality brands (low-quality brands) led to a sensible nested market structure. Stated otherwise, the choice task designs were created in such a way that the highest price level was never shown with a low-quality brand and, vice versa, the lowest price level never appeared with a high-quality brand. Similarly, for the treatments with five price levels (see Table 1, factor 3, A5L5 and A9L5), the choice tasks were generated in a way that only the highest, second highest, and third highest price levels were allowed for high-quality brands (with incidence rates of 60%, 30%, and 10%, respectively), and only the lowest, second lowest, and third lowest price levels could occur together with the low-quality brands (again with incidence rates of 60%, 30%, and 10%, respectively). Therefore, only the medium price level could appear with both subsets of brands.

Part-worth utilities: As described in Sect. 2.2, different nest-specific utility ranges (importances) for the brand and price attributes were generated to accommodate the nested structure, i.e. that the brand (as proxy for quality) receives much more importance than the price in the high-quality tier, while the reverse in the low-quality nest. Specifically, the (mean) part-worths for the price attribute were firstly order-constrained subsequent to their random generation so that higher price levels were less preferred. In order to assign the generated "true" mean part-worth utilities for the brand and price attribute to the high- and low-quality nests the part-worth utilities were sorted. In the treatments with two (three) brands per nest, the two (three) largest brand constants were assigned to the high-quality nest, and the two (three) lowest brand constants to the low-quality nest. Similarly, according to the incidence rates for the price levels (see Sect. 2.2), the price part-worths were assigned to the two nests. Finally, for model estimation, the medium price level, the least preferred brand and all first levels of the remaining attributes were chosen to constitute the reference category, i.e. the corresponding mean part-worths were set to zero by shifting the part-worths appropriately.

Based on the generated mean part-worths, individual-level part-worth utilities were generated in the nests by adding a heterogeneity component. For this, we assumed that the nest-specific part-worth utilities follow a multivariate normal distribution with mean vector b_{TRUE} and covariance matrix W_{TRUE}, where the covariance matrix W_{TRUE} captures the amount of heterogeneity across respondents, respectively. Therefore, N_1 and N_2 ($N_2 = 1{,}000 - N_1$) quantities were drawn from a normal distribution with mean 0 and a variance of either 0.05 (low heterogeneity) or 0.25 (high heterogeneity) and added to the mean part-worth utilities. Adding random error terms to the price part-worths was controlled to preserve the order constraint that higher price levels should have assigned lower utility values at the individual respondent level, too. Hence, random draws that led to price reversals at the individual respondent level were rejected. Prohibiting price reversals eliminates one possible confounding factor for our model comparison between the HB-MNL and HB-NMNL model.

Simulation of choices: Simulating choices for respondents under a MNL model is standard and proceeds as follows: iid Gumbel-distributed error terms are generated and added to obtain stochastic utilities $U_{njs} = V_{njs} + \varepsilon_{njs}$. Subsequently, the alternative with the highest stochastic utility is assigned as the chosen alternative. Simulating choices based on error components that replicate the assumed correlation of a nested logit structure is not such straightforward, as error terms are identically but not independently distributed. Garrow et al. (2010) compared three methods to determine simulated choices of respondents for the NMNL model: (1) "roulette wheel selection" (Lipowski and Lipowska 2012) based on nested logit probabilities, (2) creating error terms based on a normal distribution to approximate a multivariate Gumbel distribution and adding them to the deterministic utilities to obtain stochastic utilities, and (3) generating bivariate Gumbel errors using the Type C bivariate Gumbel distribution and adding them to the deterministic utilities to obtain stochastic utilities. The findings of Garrow et al. (2010) clearly indicated that the first method that uses closed-form nested logit probabilities should be used, as it provided an excellent and unbiased recovery of parameters of nested structures. Therefore, we relied on this method to simulate respondents' choices in our Monte Carlo study. The method works as follows: assume that for three alternatives offered in a choice set the nested logit probabilities $\mathrm{P}_{\mathrm{njs}}^{\mathrm{NMNL}}$ are $\{0.3, 0.6, 0.1\}$. Then, standard uniform draws with values less than 0.3 result in alternative one being chosen, with values between 0.3 and 0.9 in alternative two being chosen, and with values greater than 0.9 in alternative three being chosen. Note that the nested logit probabilities depend on the log-sum coefficients given under a respective treatment, compare Eq. (2) in Sect. 2.1.

Appendix 3: Additional Tables

See Tables 6, 7, 8, 9.

Table 6 F-Tests[a] of main and interaction effects on parameter recovery, goodness-of-fit and predictive accuracy (N = 1280; p-values in parentheses)

Source (degrees of freedom)	Parameter recovery			Goodness-of-fit				Predictive accuracy				
	Correlation	RMSE	MAE	LL	PC	BS	SpS	Hit rate	BS	SpS	LL	RMSE(V)
Model (1)	**14.947 (<0.001)**	**141.192 (<0.001)**	**182.395 (<0.001)**	**53.948 (<0.001)**	**102.348 (<0.001)**	**39.526 (<0.001)**	**45.093 (<0.001)**	0.730 (0.393)	0.092 (0.761)	0.029 (0.866)	**11.986 (<0.001)**	**73.675 (<0.001)**
Nest size (4)	1.411 (0.228)	**5.986 (<0.001)**	**6.028 (<0.001)**	**10.370 (<0.001)**	**9.178 (<0.001)**	**8.407 (<0.001)**	**10.706 (<0.001)**	**16.921 (<0.001)**	**14.100 (<0.001)**	**15.184 (<0.001)**	**11.862 (<0.001)**	**4.443 (0.001)**
Brands (1)	0.449 (0.503)	**30.120 (<0.001)**	**37.964 (<0.001)**	**265.244 (<0.001)**	**7.748 (0.005)**	**169.348 (<0.001)**	**191.409 (<0.001)**	**157.133 (<0.001)**	**302.574 (<0.001)**	**196.261 (<0.001)**	**380.419 (<0.001)**	**10.757 (0.001)**
Complexity (3)	**46.091 (<0.001)**	**40.839 (<0.001)**	**79.678 (<0.001)**	**10.000 (<0.001)**	**33.610 (<0.001)**	**12.214 (<0.001)**	**15.858 (<0.001)**	**14.694 (<0.001)**	**15.106 (<0.001)**	**15.640 (<0.001)**	**24.991 (<0.001)**	**2.725 (0.043)**
Similarity (3)	**7.674 (<0.001)**	**42.780 (<0.001)**	**59.478 (<0.001)**	**245.688 (<0.001)**	**226.288 (<0.001)**	**238.808 (<0.001)**	**273.759 (<0.001)**	**63.494 (<0.001)**	**51.756 (<0.001)**	**57.569 (<0.001)**	**33.894 (<0.001)**	**23.399 (<0.001)**
Heterogeneity (1)	**51.971 (<0.001)**	**138.352 (<0.001)**	**77.350 (<0.001)**	**29.014 (<0.001)**	**26.767 (<0.001)**	**11.961 (<0.001)**	**6.871 (0.009)**	**275.061 (<0.001)**	**270.346 (<0.001)**	**297.225 (<0.001)**	**312.057 (<0.001)**	**73.778 (<0.001)**
Model × Heterogeneity	1.355 (0.245)	**11.374 (<0.001)**	**9.529 (0.002)**	3.013 (0.083)	3.203 (0.074)	2.869 (0.091)	3.551 (0.060)	0.015 (0.902)	0.056 (0.813)	0.012 (0.913)	0.174 (0.677)	**6.775 (0.009)**
Model × Similarity	**4.989 (0.002)**	**44.149 (<0.001)**	**64.724 (<0.001)**	**12.415 (<0.001)**	**8.963 (<0.001)**	**10.871 (<0.001)**	**12.424 (<0.001)**	0.090 (0.966)	0.039 (0.990)	0.001 (1.000)	1.970 (0.117)	**22.815 (<0.001)**
Model × Complexity	0.599 (0.616)	**9.465 (<0.001)**	**18.745 (<0.001)**	0.248 (0.863)	**6.052 (<0.001)**	0.221 (0.882)	0.240 (0.868)	0.007 (0.999)	0.027 (0.994)	0.019 (0.996)	0.418 (0.740)	2.451 (0.062)
Model × Brands	1.223 (0.269)	**7.576 (0.006)**	**8.902 (0.003)**	2.681 (0.102)	**4.762 (0.029)**	0.012 (0.912)	2.203 (0.138)	0.024 (0.878)	0.001 (0.974)	0.008 (0.930)	0.002 (0.963)	3.277 (0.071)
Model × Nest size	0.240 (0.916)	0.339 (0.852)	1.001 (0.406)	0.122 (0.974)	0.153 (0.962)	0.170 (0.954)	0.155 (0.961)	0.001 (1.000)	0.001 (1.000)	0.000 (1.000)	0.023 (0.999)	0.535 (0.710)
Heterogeneity × Similarity	**5.577 (<0.001)**	**4.180 (0.006)**	**6.460 (<0.001)**	**9.521 (<0.001)**	**8.336 (<0.001)**	**8.793 (<0.001)**	**9.663 (<0.001)**	**13.067 (<0.001)**	**11.416 (<0.001)**	**12.461 (<0.001)**	**10.183 (<0.001)**	2.233 (0.083)
Heterogeneity × Complexity	**3.512 (0.015)**	0.946 (0.418)	2.042 (0.106)	2.508 (0.057)	1.907 (0.127)	1.540 (0.202)	1.552 (0.200)	**4.011 (0.007)**	**3.686 (0.012)**	**4.149 (0.006)**	**4.859 (0.002)**	1.316 (0.268)
Heterogeneity × Brands	**17.971 (<0.001)**	**13.401 (<0.001)**	**11.891 (<0.001)**	1.135 (0.287)	0.276 (0.599)	0.018 (0.895)	0.117 (0.732)	**6.437 (0.011)**	0.679 (0.410)	**9.682 (0.002)**	**20.377 (<0.001)**	**8.011 (0.005)**

Table 6 (continued)

Source (degrees of freedom)	Parameter recovery			Goodness-of-fit				Predictive accuracy				
	Correlation	RMSE	MAE	LL	PC	BS	SpS	Hit rate	BS	SpS	LL	RMSE(V)
Heterogeneity × Nest size	**26.908**	1.283	1.553	**9.357**	**9.144**	**9.653**	**10.303**	**9.539**	**9.746**	**10.080**	**9.652**	1.131
	(<0.001)	(0.275)	(0.185)	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	(0.340)
Similarity × Complexity	**2.015**	**3.048**	**5.129**	**2.006**	**2.700**	1.462	**2.122**	0.608	0.498	0.644	1.047	**2.599**
	(0.035)	**(0.001)**	**(<0.001)**	**(0.035)**	**(0.004)**	(0.157)	**(0.025)**	(0.791)	(0.877)	(0.760)	(0.400)	**(0.006)**
Similarity × Brands	1.345	0.978	1.741	**8.579**	2.168	1.036	**8.735**	0.877	**4.114**	0.876	1.228	0.245
	(0.258)	(0.402)	(0.157)	**(<0.001)**	(0.090)	(0.376)	**(<0.001)**	(0.452)	**(0.006)**	(0.453)	(0.298)	(0.865)
Similarity × Nest size	1.568	**5.592**	**6.212**	**11.927**	**10.992**	**13.738**	**15.262**	**5.154**	**3.931**	**4.357**	**3.186**	**3.500**
	(0.095)	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**
Complexity × Brands	0.905	**3.310**	**3.512**	1.639	**15.020**	1.835	1.551	2.061	2.118	2.006	1.124	**4.241**
	(0.438)	**(0.020)**	**(0.015)**	(0.179)	**(<0.001)**	(0.139)	(0.200)	(0.104)	(0.096)	(0.111)	(0.338)	**(0.005)**
Complexity × Nest size	**2.441**	**2.055**	1.709	**5.290**	**5.201**	**4.670**	**4.880**	**6.043**	**4.881**	**5.931**	**5.982**	**2.229**
	(0.004)	**(0.017)**	(0.060)	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(<0.001)**	**(0.009)**
Brands × Nest size	0.371	1.504	0.852	**3.203**	2.197	1.359	**3.208**	**2.602**	1.287	**2.770**	**2.949**	0.994
	(0.829)	(0.199)	(0.493)	**(0.013)**	(0.067)	(0.246)	**(0.012)**	**(0.035)**	(0.273)	**(0.026)**	**(0.019)**	(0.409)
R^2	0.287	0.439	0.503	0.563	0.530	0.532	0.568	0.452	0.467	0.460	0.495	0.284
(Adjusted R^2)	(0.240)	(0.402)	(0.471)	(0.534)	(0.499)	(0.501)	(0.540)	(0.416)	(0.431)	(0.425)	(0.462)	(0.236)

[a]Bold values indicate significant differences at the 0.05 level

Table 7 Interaction effects between number of respondents per nest (nest size) and degree of similarity between alternatives on model fit measures[a]

	Goodness-of-fit			
	Brier score			
Nest size/similarity	(1) 0202	(2) 0208	(3) 0802	(4) 0808
(1) 125 875	−0.036	−0.060^{3,4,5*}	−0.039^{2,3,4,5*}	−0.070^{5*}
(2) 250 750	−0.035	−0.063^{3,4,5*}	−0.051^{1*}	−0.067
(3) 500 500	−0.036	−0.049^{1,2*}	−0.048^{1*}	−0.066
(4) 750 250	−0.037	−0.044^{1,2*}	−0.054^{1*}	−0.063
(5) 875 125	−0.031	−0.041^{1,2*}	−0.056^{1*}	−0.061^{1*}
	Spherical score			
Nest size/Similarity	(1) 0202	(2) 0208	(3) 0802	(4) 0808
(1) 125 875	0.902	0.832^{3,4,5*}	0.892^{2,3,4,5*}	0.800^{5*}
(2) 250 750	0.902	0.822^{3,4,5*}	0.857^{1*}	0.809
(3) 500 500	0.901	0.864^{1,2,5*}	0.867^{1*}	0.807^{5*}
(4) 750 250	0.900	0.879^{1,2*}	0.849^{1*}	0.822
(5) 875 125	0.913	0.887^{1,2,3*}	0.844^{1*}	0.833^{1,3*}

[a]Superscripts on means refer to the factor levels and indicate significant differences at the 0.05 level (considering Bonferroni correction). Underlined values denote the factor levels showing the best performance w.r.t. the measure considered, respectively

Table 8 Interaction effects between type of model and degree of similarity between alternatives in the nests on parameter recovery measured in absolute errors (RMSE and MAE)[a]

		Parameter recovery	
Model	Similarity	RMSE	MAE
NMNL	(1) 0202	0.834	0.542
	(2) 0208	0.834	0.562
	(3) 0802	0.804	0.539
	(4) 0808	0.831	0.546
MNL	(1) 0202	1.498^{2,3,4*}	0.990^{2,3,4*}
	(2) 0208	0.983^{1,3,4*}	0.644^{1,3,4*}
	(3) 0802	1.230^{1,2,4*}	0.807^{1,2,4*}
	(4) 0808	0.751^{1,2,3*}	0.488^{1,2,3*}

[a]Superscripts on means refer to the factor levels and indicate significant differences at the 0.05 level (considering Bonferroni correction). Underlined values denote the factor levels showing the best performance w.r.t. the measure considered, respectively

Table 9 Interaction effect between number of respondents per nest (nest size) and sample heterogeneity on parameter recovery measured by mean Pearson correlations[a]

	Parameter recovery	
	Mean Pearson correlation	
Nest size/heterogeneity	(1) 005	(2) 025
(1) 125 875	0.969^{4,5*}	0.938^{3,4,5*}
(2) 250 750	0.971^{4,5*}	0.939^{3,4,5*}
(3) 500 500	0.960^{5*}	0.948^{1,2,4*}
(4) 750 250	0.950^{1,2*}	0.958^{1,2,3*}
(5) 875 125	0.946^{1,2*}	0.953^{1,2*}

[a]Superscripts on means refer to the factor levels and indicate significant differences at the 0.05 level (considering Bonferroni correction). Underlined values denote the factor levels showing the best performance w.r.t. the measure considered, respectively

Funding Open Access funding enabled and organized by Projekt DEAL. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability All code used for data generation, estimation, and validation − comprising R and SAS scripts − is publicly accessible via the Open Science Framework (OSF) at https://osf.io/9afxc/.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Abramson C, Andrews RL, Currim IS, Jones M (2000) Parameter bias from unobserved effects in the multinomial logit model of consumer choice. J Mark Res 37(4):410–426. https://doi.org/10.1509/jmkr.37.4.410.18791

Ailawadi KL, Gedenk K, Lutzky C, Neslin SA (2007) Decomposition of the sales impact of promotion-induced stockpiling. J Mark Res 44(3):450–467. https://doi.org/10.1509/jmkr.44.3.450

Akinc D, Vandebroek M (2018) Bayesian estimation of mixed logit models: selecting an appropriate prior for the covariance matrix. J Choice Model 29:133–151. https://doi.org/10.1016/j.jocm.2017.11.004

Allenby GM, Ginter JL (1995) Using extremes to design products and segment markets. J Mark Res 32(4):392. https://doi.org/10.2307/3152175

Allenby GM, Rossi PE (1991) Quality perceptions and asymmetric switching between brands. Mark Sci 10(3):185–204. https://doi.org/10.1287/mksc.10.3.185

Allenby GM, Rossi PE (1998) Marketing models of consumer heterogeneity. J Econom 89(1–2):57–78. https://doi.org/10.1016/S0304-4076(98)00055-4

Allenby GM, Arora N, Ginter JL (1995) Incorporating prior knowledge into the analysis of conjoint studies. J Mark Res 32(2):152–162. https://doi.org/10.2307/3152044

Allenby GM, Arora N, Ginter JL (1998) On the heterogeneity of demand. J Mark Res 35(3):384–389. https://doi.org/10.2307/3152035

Amemiya T, Shimono K (1989) An application of nested logit models to the labor supply of the elderly. Econ Stud Q 40(1):14–22

Anderson SP, Palma AD (1992) Multiproduct firms: a nested logit approach. J Ind Econ 40(3):261–276. https://doi.org/10.2307/2950539

Andrews RL, Ainslie A, Currim IS (2002a) An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. J Mark Res 39(4):479–487. https://doi.org/10.1509/jmkr.39.4.479.19124

Andrews RL, Ansari A, Currim IS (2002b) Hierarchical Bayes versus finite mixture v analysis models: a comparison of fit, prediction, and partworth recovery. J Mark Res 39(1):87–98. https://doi.org/10.1509/jmkr.39.1.87.18936

Aribarg A, Burson KA, Larrick RP (2017) Tipping the scale: the role of discriminability in conjoint analysis. J Mark Res 54(2):279–292. https://doi.org/10.1509/jmr.14.0659

Baltas G, Doyle P, Dyson P (1997) A model of consumer choice for national vs private label brands. J Oper Res Soc 48(10):988–995. https://doi.org/10.1057/palgrave.jors.2600454

Belloni A, Freund R, Selove M, Simester D (2008) Optimizing product line designs: efficient methods and comparisons. Manag Sci 54(9):1544–1552. https://doi.org/10.1287/mnsc.1080.0864

Ben-Akiva ME (1979) Disaggregate travel and mobility choice models and measures of accessibility. In: Stopher PR (ed) Hensher DA. Behavioural Travel Modelling, Croom Helm, pp 654–679

Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis: Theory and application to travel demand. MIT Press

Ben-Akiva ME (1973) Structure of passenger travel demand models. Dissertation, Massachusetts Institute of Technology

Berkovec J, Rust J (1985) A nested logit model of automobile holdings for one vehicle households. Transp Res Part B Methodol 19(4):275–285. https://doi.org/10.1016/0191-2615(85)90036-0

Berry ST (1994) Estimating discrete-choice models of product differentiation. RAND J Econ 25(2):242–262. https://doi.org/10.2307/2555829

Berry ST, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. Econometrica 63:841–890. https://doi.org/10.2307/2171802

Bhat CR (1997) Covariance heterogeneity in nested logit models: econometric structure and application to intercity travel. Transp Res Part B Methodol 31(1):11–21. https://doi.org/10.1016/S0191-2615(96)00018-5

Blattberg RC, Wisniewski KJ (1989) Price-induced patterns of competition. Mark Sci 8(4):291–309. https://doi.org/10.1287/mksc.8.4.291

Blattberg RC, Briesch R, Fox EJ (1995) How promotions work. Mark Sci 14(3_supplement):G122–G132. https://doi.org/10.1287/mksc.14.3.G122

Boxebeld S (2024) Ordering effects in discrete choice experiments: a systematic literature review across domains. J Choice Model 51:100489. https://doi.org/10.1016/j.jocm.2024.100489

Boyd JH, Mellman RE (1980) The effect of fuel economy standards on the U.S. automotive market: an hedonic demand analysis. Transp Res Part A Gen 14(5–6):367–378. https://doi.org/10.1016/0191-2607(80)90055-2

Braun A, Schmeiser H, Schreiber F (2016) On consumer preferences and the willingness to pay for term life insurance. Eur J Oper Res 253(3):761–776. https://doi.org/10.1016/j.ejor.2016.02.023

Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78(1):1–3. https://doi.org/10.1175/1520-0493(1950)078%3c0001:VOFEIT%3e2.0.CO;2

Cardell NS (1997) Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. Econometr Theory 13(2):185–213. https://doi.org/10.1017/s0266466600005727

Cardell NS, Dunbar FC (1980) Measuring the societal impacts of automobile downsizing. Transp Res Part A Gen 14(5–6):423–434. https://doi.org/10.1016/0191-2607(80)90060-6

Carlsson F, Mørkbak MR, Olsen SB (2012) The first time is the hardest: a test of ordering effects in choice experiments. J Choice Model 5(2):19–37

Chatterjee P, Kumar A (2017) Consumer willingness to pay across retail channels. J Retail Consumer Serv 34:264–270. https://doi.org/10.1016/j.jretconser.2016.01.008

Chib S, Seetharaman PB, Strijnev A (2004) Model of brand choice with a nopurchase option calibrated to scanner-panel data. J Mark Res 41(2):184–196. https://doi.org/10.1509/jmkr.41.2.184.28674

Chintagunta PK, Vilcassim NJ (1998) Empirical implications of unobserved household heterogeneity for manufacturer and retailer pricing. J Retail Consumer Serv 5(1):15–24. https://doi.org/10.1016/S0969-6989(96)00061-6

Cohen J (1988) Statistical power analysis for the behavioral sciences (2nd ed). Lawrence Erlbaum Associates

Crawford GS, Griffith R, Iaria A (2021) A survey of preference estimation with unobserved choice set heterogeneity. J Econom 222(1):4–43. https://doi.org/10.1016/j.jeconom.2020.07.024

Daly A, Zachary S (1978) Improved multiple choice models. In: Dalvi MQ (ed) Hensher DA. Determinants of travel choice, Saxon House, pp 335–357

Day B, Prades JLP (2010) Ordering anomalies in choice experiments. J Environ Econ Manag 59:271–285. https://doi.org/10.1016/j.jeem.2010.03.001

Day B, Bateman I, Carson R, Dupont D, Louviere J, Morimoto S, Scarpa R, Wang P (2012) Ordering effects and choice set awareness in repeat-response stated preference studies. J Environ Econ Manag 63(1):73–91. https://doi.org/10.1016/j.jeem.2011.09.001

Debreu G (1960) Review of RD luce individual choice behavior. Am Econ Rev 50(1):186–188. https://doi.org/10.1090/S0002-9904-1960-10452-1

Dumont J, Keller J, Carpenter C (2019) Functions for hierarchical bayesian estimation: a flexible approach. https://doi.org/10.32614/cran.package.rsghb

Elrod T, Kumar SK (1989) Bias in the first choice rule for predicting share. In: Proceedings of the 1989 Sawtooth software conference, pp 259–271

Elshiewy O, Guhl D, Boztuğ Y (2017) Multinomial logit models in marketing—from fundamentals to state-of-the-art. Mark ZFP 39(3):32–49. https://doi.org/10.15358/0344-1369-2017-3-32

Falaris EM (1987) A nested logit migration model with selectivity. Int Econ Rev 28(2):429–443. https://doi.org/10.2307/2526735

Gabor A, Granger CWJ (1966) Price as an indicator of quality: report on an enquiry. Economica 33(129):43–70. https://doi.org/10.2307/2552272

Garrow LA, Bodea TD, Lee M (2010) Generation of synthetic datasets for discrete choice analysis. Transp 37(2):183–202. https://doi.org/10.1007/s11116-009-9228-6

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7(4):457–472. https://doi.org/10.1214/ss/1177011136

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014) Bayesian data analysis, 3rd edn. Chapman & Hall/CRC, New York

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Gensler S, Hinz O, Skiera B, Theysohn S (2012) Willingness-to-pay estimation with choice-based conjoint analysis: addressing extreme response behavior with individually adapted designs. Eur J Oper Res 219(2):368–378. https://doi.org/10.1016/j.ejor.2012.01.002

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437

Goeken N, Kurz P, Steiner WJ (2021) Hierarchical Bayes conjoint choice models – model framework, Bayesian inference, model selection, and interpretation of estimation results. Mark ZFP 43(3):49–64. https://doi.org/10.15358/0344-1369-2021-3-49

Goeken N, Kurz P, Steiner W (2024) Multimodal preference heterogeneity in choice-based conjoint analysis: a simulation study. J Bus Econ 94:137–185. https://doi.org/10.1007/s11573-023-01156-6

Grigolon L, Verboven F (2014) Nested logit or random coefficients logit? A comparison of alternative discrete choice models of product differentiation. Rev Econ Stat 96(5):916–935. https://doi.org/10.1162/rest_a_00420

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1):97–109. https://doi.org/10.2307/2334940

Hauser JR (1978) Testing the accuracy, usefulness, and significance of probabilistic choice models: an information-theoretic approach. Oper Res 26(3):406–421. https://doi.org/10.1287/opre.26.3.406

Hein M, Kurz P, Steiner WJ (2019a) Analyzing the capabilities of the HB logit model for choice-based conjoint analysis: a simulation study. J Bus Econ 90:1–36. https://doi.org/10.1007/s11573-019-00927-4

Hein M, Kurz P, Steiner WJ (2019b) On the effect of HB covariance matrix prior settings: a simulation study. J Choice Model 31:51–72. https://doi.org/10.1016/j.jocm.2019.02.001

Hess S, Hensher D, Daly A (2012) Not bored yet—revisiting respondent fatigue in stated choice experiments. Transp Res Part A Pol Prac 46(3):626–644. https://doi.org/10.1016/j.tra.2011.11.008

Hoffman SD, Duncan GJ (1988) A comparison of choice-based multinomial and nested logit models: the family structure and welfare use decisions of divorced or separated women. J Hum Resour 23(4):550–562. https://doi.org/10.2307/145813

Hoogerbrugge M, van der Wagt K (2006) How many choice tasks should we ask? In: Proceedings of the 2006 sawtooth software conference, pp 97–110

Van Horn K, McDonnel Feit E (2017) Stan tutorial: module 4: hierarchical nested multinomial logit. ART-Forum 2017. https://github.com/ksvanhorn/ART-Forum-2017-Stan-Tutorial/blob/master/ 4_hnestedmnl/hnestedmnl.pdf

Huber J, Meissner M, Musalem A (2018) How does choice-based conjoint work? In Proceedings of the 20th sawtooth software conference, Orlando

Huber, J (1998) Achieving individual-level predictions from CBC data: comparing ICE and hierarchical Bayes. Sawtooth Software Research Paper Series

Karniouchina EV, Moore WL, van der Rhee B, Verma R (2009) Issues in the use of ratings-based versus choice-based conjoint analysis in operations management research. Eur J Oper Res 197(1):340–348. https://doi.org/10.1016/j.ejor.2008.05.029

Kopalle PK, Kannan PK, Boldt LB, Arora N (2012) The impact of household level heterogeneity in reference price effects on optimal retailer pricing policies. J Retail 88(1):102–114. https://doi.org/10.1016/j.jretai.2011.10.002

Kuhfeld WF, Tobias RD, Garratt M (1994) Efficient experimental design with marketing research applications. J Mark Res 31(4):545–557. https://doi.org/10.2307/3151882

Kuhfeld WF (1997) Efficient experimental designs using computerized searches. Sawtooth software research paper series

Kurz P, Binner S (2012) The individual choice task threshold: need for variable number of choice tasks. In Proceedings of the 2012 sawtooth software conference, pp 111–127

Lahiri K, Gao J (2002) Bayesian analysis of nested logit model by markov chain monte carlo. J Econ 111(1):103–133. https://doi.org/10.1016/S0304-4076(02)00125-2

Lee B (1999) Calling patterns and usage of residential toll service under self selecting tariffs. J Regul Econ 16(1):45–82. https://doi.org/10.1023/A:1008131714712

Lenk PJ, DeSarbo WS, Green PE, Young MR (1996) Hierarchical bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs. Mark Sci 15(2):173–191. https://doi.org/10.1287/mksc.15.2.173

Lipowski A, Lipowska D (2012) Roulette-wheel selection via stochastic acceptance. Physica A Stat Mech Appl 391(6):2193–2196. https://doi.org/10.1016/j.physa.2011.12.004

Louviere JJ, Woodworth G (1983) Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. J Mark Res 20(4):350–367. https://doi.org/10.2307/3151440

Louviere JJ, Flynn TN, Carson RT (2010) Discrete choice experiments are not conjoint analysis. J Choice Model 3(3):57–72. https://doi.org/10.1016/S1755-5345(13)70014-9

McFadden D (1977) Quantitative methods for analyzing travel behavior of individuals: some recent developments. Cowles Foundation Discussion Papers 474. Cowles Foundation for Research in Economics, Yale University

McFadden D (1978) Modeling the choice of residential location. In: A Karlqvist, L Lundqvist, F Snickars, J Weibull (Eds) Spatial interaction theory and planning models, North-Holland Publishing Company, pp 75–96

McFadden D (1981) Econometric models of probabilistic choice. In: Manski CF, McFadden D (Eds) Structural analysis of discrete data with econometric applications, MIT Press, pp 198–272

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092. https://doi.org/10.1063/1.1699114

Moore WL (2004) A cross-validity comparison of rating-based and choice-based conjoint analysis models. Int J Res Mark 21(3):299–312. https://doi.org/10.1016/j.ijresmar.2004.01.002

Moore WL, Gray-Lee J, Louviere JJ (1998) A cross-validity comparison of conjoint analysis and choice models at different levels of aggregation. Mark Lett 9(2):195–207. https://doi.org/10.1023/A:1007913100332

Natter M, Feurstein M (2002) Real world performance of choice-based conjoint models. Eur J Oper Res 137(2):448–458. https://doi.org/10.1016/S0377-2217(01)00147-3

Nguyen TC, Le HT, Nguyen HD, Ngo MT, Nguyen HQ (2021) Examining ordering effects and strategic behaviour in a discrete choice experiment. Econ Anal Policy 70:394–413. https://doi.org/10.1016/j.eap.2021.03.005

Ogawa K (1987) An approach to simultaneous estimation and segmentation in conjoint analysis. Mark Sci 6(1):66–81. https://doi.org/10.1287/mksc.6.1.66

Orme B (2024): A comparison of HB-MNL estimation via Metropolis Hastings (Sawtooth Software's CBC/HB Program), Hamiltonian Monte Carlo (via Stan), and Variational Bayes (via Stan). Sawtooth Software Research Paper Series, Utah

Otter T, Tüchler R, Frühwirth-Schnatter S (2004) Capturing consumer heterogeneity in metric conjoint analysis using Bayesian mixture models. Int J Res Mark 21(3):285–297. https://doi.org/10.1016/j.ijresmar.2003.11.002

Paetz F, Steiner WJ (2017) The benefits of incorporating utility dependencies in finite mixture probit models. Or Spectr 39(3):793–819. https://doi.org/10.1007/s00291-017-0478-y

Poirier DJ (1996) A bayesian analysis of nested logit models. J Econom 75(1):163–181. https://doi.org/10.1016/0304-4076(95)01775-5

Rao AR, Monroe KB (1989) The effect of price, brand name, and store name on buyers' perceptions of product quality: an integrative review. J Mark Res 26(3):351–357. https://doi.org/10.2307/3172907

Rossi PE, Allenby GM (2000) Statistics and marketing. J Am Stat Assoc 95(450):635–638. https://doi.org/10.2307/2669407

Savage S, Waldman D (2008) Learning and fatigue during choice experiments: a comparison of online and mail survey modes. J Appl Econ 23(3):351–371. https://doi.org/10.1002/jae.984

Silberhorn N, Boztuğ Y, Hildebrandt L (2008) Estimation with the nested logit model: specifications and software particularities. Or Spectr 30(4):635–653. https://doi.org/10.1007/s00291-007-0109-0

Sivakumar K (1995) Role of price and quality tiers on the cluster effect in brand choice. Mark Lett 6(4):265–273. https://doi.org/10.1007/BF00996190

Sawtooth Software (2016) Software for Hierarchical Bayes Estimation for CBC Data, CBC/HB v5. Sawtooth Software, Inc.

Train KE (1980) A structured logit model of auto ownership and mode choice. Rev Econ Stud 47(2):357–370. https://doi.org/10.2307/2296997

Train KE, McFadden D, Ben-Akiva ME (1987) The demand for local telephone service: a fully discrete model of residential calling patterns and service choices. RAND J Econ 18(1):109–123. https://doi.org/10.2307/2555538

Train KE, Ben-Akiva ME, Atherton T (1989) Consumption patterns and self-selecting tariffs. Rev Econ Stat 71(1):62–73. https://doi.org/10.2307/1928052

Train KE, Sonnier G (2005) Mixed logit with bounded distributions of correlated partworths. In: Scarpa R, Alberini A (Eds.) Applications of simulation methods in environmental and resource economics. The economics of non-market goods and resources, Springer, pp 117–134. https://doi.org/10.1007/1-4020-3684-1_7

Train KE (2001) A comparison of hierarchical Bayes and maximum simulated likelihood for mixed logit. Working Paper, Department of Economics, University of California

Train KE (2009) Discrete choice methods with simulation (2nd ed). Cambridge University Press

Van Horn K (2024) Comparison of algorithms for estimating Bayesian HMNL: MH vs. Unpublished white paper, HMC vs. VB

Verlegh PW, Schifferstein HN, Wittink DR (2002) Range and number-of-levels effects in derived and stated measures of attribute importance. Mark Lett 13(1):41–52. https://doi.org/10.1023/A:1015063125062

Voleti S, Srinivasan V, Ghosh P (2017) An approach to improve the predictive power of choice-based conjoint analysis. Int J Res Mark 34(2):325–335. https://doi.org/10.1016/j.ijresmar.2016.08.007

Vriens M, Wedel M, Wilms T (1996) Metric conjoint segmentation methods: a monte carlo comparison. J Mark Res 33(1):73–85. https://doi.org/10.1177/002224379603300107

Williams HCWL (1977) On the formation of travel demand models and economic evaluation measures of user benefit. Environ Plan a: Econ Space 9(3):285–344. https://doi.org/10.1068/a090285

Wirth R (2010) HB-CBC, HB-best-worst-CBC or no HB at all? Proceedings of the 2010 sawtooth software conference, pp 321–355

Zwerina K, Huber J, Kuhfeld WF (1996) A general method for constructing efficient choice designs. Working Paper, Fuqua School of Business, Duke University

Authors and Affiliations

Nils Goeken[1] · **Peter Kurz**[2] · **Winfried J. Steiner**[1]

✉ Nils Goeken
nils.goeken@tu-clausthal.de

✉ Winfried J. Steiner
winfried.steiner@tu-clausthal.de

Peter Kurz
p.kurz@bms-net.de

1 Department of Marketing, Clausthal University of Technology, 38678 Clausthal-Zellerfeld, Germany

2 bms marketing research + strategy, Landsberger Straße 487, 81241 Munich, Germany