European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Using Hierarchical Bayes draws for improving shares of choice predictions in conjoint simulations: A study based on conjoint choice data

Maren Hein[a], Nils Goeken[b], Peter Kurz[c], Winfried J. Steiner[b,*]

[a] State Office for Statistics Lower Saxony, Göttinger Chaussee 76, 30453 Hannover, Germany
[b] Department of Marketing, Clausthal University of Technology, 38678 Clausthal-Zellerfeld, Germany
[c] bms marketing research + strategy, Landsberger Straße 487, 81241 München, Germany

ARTICLE INFO

Article history:
Received 3 December 2018
Accepted 31 May 2021
Available online xxx

Keywords:
OR in marketing
Choice rules
Choice-based conjoint analysis
Preference shares
IIA property

ABSTRACT

The use of Hierarchical Bayes (HB) estimation techniques for choice-based conjoint (CBC) data offers the opportunity to directly use HB draws for preference simulations. This paper analyzes the use of HB draws for shares of choice predictions. Five different choice rules are compared: the first choice rule applied to HB draws, the logit choice rule applied to HB draws, the randomized first choice rule, the traditional first choice rule and the traditional logit choice rule. Each two different holdout choice scenarios are constructed containing one time two extremely similar and the other time very unique alternatives to assess how well the choice rules tolerate the IIA property in predicting choice shares. We present a Monte Carlo study to systematically explore shares of choice predictions based on the different choice rules and further verify whether our findings hold in empirical settings. The key finding of our Monte Carlo study is that using HB draws either combined with the first choice rule or the logit choice rule substantially improves choice share predictions when compared to the other choice rules, regardless of the type of holdout choice scenario. While the logit choice rule applied to HB draws performs a touch better for simulated data, the first choice rule applied to HB draws provides the best choice share predictions for each of the five empirical data sets. Using HB draws does not only provide the best predictive validity but, more importantly, it is theoretically correct when applying a Bayesian estimation approach to CBC data.

1. Introduction

Since its introduction to marketing research (Green & Rao, 1971), conjoint analysis has become a widely applied method for measuring, analyzing, and predicting consumer preferences. In recent years, (choice-based) conjoint analysis has also expanded to other disciplines with increasing intensity in operations management contexts (Braun, Schmeiser & Schreiber, 2016; Camm, Cochran, Curry & Kannan, 2006; Foster, Turner & Ferguson, 2014; Gensler, Hinz, Skiera & Theysohn, 2012; Halme & Kallio, 2011, 2014; Karniouchina, Moore, van der Rhee & Verma, 2009; Maldonado, Montoya & Weber, 2015; Meeran, Jahanbin, Goodwin & Neto, 2016; Natter & Feurstein, 2002; Steiner, 2010; Tsafarakis, Grigoroudis & Matsatsinis, 2011; Wang, Camm & Curry, 2009). One of the main reasons for applying conjoint analysis in marketing practice is to conduct preference or market simulations (e.g., Green & Srinivasan, 1990; Rao, 2014; Wittink & Cattin, 1989; Wittink, Vriens & Burhenne, 1994). Based on a (hypothetical) market scenario, preference simulations enable managers to explore how respondents may react to new product entries or product modifications among a competitive environment ("what if" scenarios), as well as predict preference shares, which are also referred to as *shares of choice* (Huber, Orme & Miller, 1999, 2007; Braun et al., 2016; Voleti, Srinivasan & Ghosh, 2017).[1] Green and Krieger (1988) recognized early the importance of simulating preference shares by stating "in virtually all industry applications of conjoint analysis, [...] individual part-worth [utility] functions, along with one or more new product descriptions, are typically entered into a buyer choice simulator to find the share of choices (i.e., "market

* Corresponding author.
E-mail addresses: maren.hein@statistik.niedersachsen.de (M. Hein), nils.goeken@tu-clausthal.de (N. Goeken), p.kurz@bms-net.de (P. Kurz), winfried.steiner@tu-clausthal.de (W. J. Steiner).

[1] In the relevant literature, the terms *shares of choice*, *shares of preference*, *choice shares*, and *preference shares* are used interchangeably, along with the terms *conjoint simulation*, *preference simulation*, and *market simulation*.

https://doi.org/10.1016/j.ejor.2021.05.056
0377-2217/

share") received by each product" (Green & Krieger, 1988, p. 114). Similarly, Natter and Feurstein (2002) noted that "in many studies, the CBC model is used to build a market simulator to develop marketing strategies; i.e., shares of preference are taken as market share forecasts" (Natter & Feurstein, 2002, p. 448). Rao (2014) argued that stable market shares provide the basis for maximizing the long-term profitability of new products or product lines (cf. Rao, 2014, p. 225). A large body of academic papers further focus on conjoint-based optimal product (line) design. These papers aim to develop efficient heuristics in NP-hard problem settings to maximize shares of choice and/or profit from product (line) sales to an aggregate of customers. An overview of this stream is provided by Belloni, Freund, Selove & Simester (2008) who compared the most important methods developed here to date. Accurately predicting preference shares is also a key element related to this issue.

Of course, care is required when interpreting simulated choice shares as expected market shares. Because preference simulations rely on several assumptions that may differ from real market conditions (e.g. equal awareness and/or availability of all products considered, no out-of-stock conditions, all relevant attributes included, no budget constraints for respondents), choice shares should be interpreted as relative *indicators* of preference rather than directly as expected market shares (Orme & Johnson, 2006).

Three steps are generally required to conduct preference simulations based on conjoint data. First, part-worth utilities previously estimated in a conjoint study have to be provided. Choice-based conjoint analysis (CBC, Louviere & Woodworth, 1983) has today become the standard approach for most conjoint applications, with the multinomial logit model (MNL) being the most widely used discrete choice model for utility estimation. In addition, incorporating heterogeneity of consumer preferences into studies has become state of the art, with part-worth utilities resulting from CBC studies (and preferably estimated on the individual respondent level using Hierarchical Bayes (HB) methods) primarily used for market simulations (Braun et al., 2016; Natter & Feurstein, 2002). Currently, the standard estimation approach for considering random taste variation in CBC studies remains the Hierarchical Bayes multinomial logit model with the multivariate normal distribution used as probability distribution, also called the HB mixed logit model (see Train, 2009, and Appendix A for more details).

Secondly, a market scenario has to be defined to explore competitive or cannibalization effects. For example, if a firm wants to modify one of its established items, the market scenario consists of the modified item plus all of the relevant existing own and competitive items (except the product prior to modification, cf. Rao, 2014). And third, to predict which items respondents would choose, choice rules are used that translate respondents' utilities into expected individual choice probabilities. These can be aggregated across respondents to obtain the share of them who prefer one product compared to the other competing items (Voleti et al., 2017).

Each choice rule has its pros and cons, and different choice rules do not necessarily lead to the same choice share predictions. The selection of the choice rule should therefore be well considered. Given that the HB mixed logit is used to estimate individual part-worth utilities (which we will assume throughout the paper), the following choice rules for predicting shares of choice are applied: (1) the first choice (FC) rule, (2) the logit choice (LC) rule, (3) the randomized first choice (RFC) rule, (4) HB random draws combined with first choice simulations (HBFC), and (5) HB random draws combined with logit choice simulations (HBLC). It is important to note that, if a Bayesian method is employed for parameter estimation (like the HB mixed logit in this study), a fully Bayesian approach must be used to derive related quantities based on this model (like shares of choice predictions in this study). This means that only the HBFC rule or the HBLC rule come into question from a theoretical perspective. Nevertheless, we can observe the usage of all five choice rules mentioned above both in the academic literature and in market research practice. The present study wants to clarify that using HB draws does not only provide superior choice share predictions compared to the other choice rules but, in particular, is theoretically correct when conducting preference simulations in a Bayesian context.

The FC rule assumes that each respondent deterministically (i.e. with certainty) chooses the product with the highest predicted utility (e.g., Elrod & Kumar, 1989; Green & Krieger, 1988; Rao, 2014). In contrast, the LC rule assumes that the probability of choosing an item is proportional to the exponential function of its expected utility, and allows "vote splitting" across items in the sense that there is a chance of choosing any of the alternatives according to their choice probability (e.g., Green & Krieger, 1988).[2] Both the FC rule and the LC rule were traditionally applied to point estimates of part-worths. But even after HB estimation techniques became available, HB draws were still averaged to obtain point estimates of part-worths for each respondent, and these point estimates were subsequently subjected to the FC and/or LC rule for shares of choice predictions (e.g., Braun et al., 2016; Foster et al., 2014; Karniouchina et al., 2009; Moore, 2004, 1998; Natter & Feurstein, 2002). The RFC rule modifies the first choice rule by repeatedly adding some random error to point estimates of part-worths (attribute variability) and/or to each total product utility (product variability). Shares of choice are then predicted by applying the first choice rule each time to the respective disturbed utilities (Huber et al., 1999; Orme & Baker, 2000; Orme & Huber, 2000). It is also possible to directly use the HB random draws generated during the HB estimation process as inputs for the first choice rule or the logit choice rule (e.g., Dotson et al., 2018; Dotson, Büschken & Allenby, 2020; Feit, Beltramo & Feinberg, 2010; Orme & Baker, 2000, Aribarg, Arora & Kang, 2010; Toubia, de Jong, Stieger & Füller, 2012) as an alternative to averaging the HB draws to point estimates in a first step and adding random error to the point estimates in a second (as with the RFC rule). Similar to the RFC rule, shares of choice predictions are produced based on hundreds or thousands of random draws instead of only single point estimates. To the best of our knowledge, only few researchers to date have utilized the estimated posterior distributions of part-worth parameters in operations research contexts, i.e. direct HB random draws for shares of choice predictions (e.g., Camm et al., 2006; Gilbride, Lenk & Brazell, 2008; Wang, Camm & Curry, 2009).

The HB mixed logit model is the standard approach for estimating decision makers' preferences based on conjoint choice data, but competing choice rules coexist in the academic literature and market research practice when it comes to predicting shares of choice ("market shares") based on these HB estimates (although the use of the posterior distribution of estimated parameters (HB draws) is conceptually the only consistent implementation within a fully Bayesian approach). This prevalence of the different choice rules raises questions about how and whether they still effectively recover preference shares. One aspect here is the degree of similarity between alternatives to which the choice share predictions apply, and it is well-known that the different choice rules are differently prone to the independence of irrelevant alternatives property (IIA) (see the detailed discussion on choice rules in Section 3). No study currently exists that has systematically analyzed the comparative performance of the different choice rules (and in particular of directly using HB draws) for shares of choice predictions. To the best of our knowledge, the comparative performance of us-

[2] This means, when the exercise of predicting shares of choice is repeated several times, each alternative is expected to be chosen with a frequency implied by its choice probability.

ing HB draws for shares of choice predictions has been assessed in only two empirical studies (Baier & Polasek, 2003; Orme & Baker, 2000), and each of these used only one single empirical data set for drawing conclusions. In addition, only the study by Orme and Baker (2000) was conducted within a discrete choice context (RFC, HBFC, HBLC), while Baier and Polasek (2003) compared the predictive performance of different choice rules (FC, LC, HBFC) for metric conjoint data. Recently, Sawtooth Software hosted a modeling competition for researchers with a CBC data set specially created for it, including 21 out-of-sample holdout tasks with different degrees of inter-product similarity (Orme, 2017). As a by-product of this modeling competition, Sawtooth Software additionally compared the performance of the HBLC rule versus the RFC rule (and also the LC rule) for shares of choice predictions.

To close this research gap, we propose a Monte Carlo study along with additional analyses using empirical data. In the Monte Carlo study, the performance of the five choice rules is evaluated under experimentally varying conditions using several statistical criteria for predictive accuracy. Analyses of variance (ANOVAs) are further conducted to evaluate the impact of the experimental factors on the measures of predictive accuracy. We further analyze how shares of choice predictions depend on the parameter recovery performance of the HB mixed logit under different experimental conditions. We also investigate the power of the different choice rules to handle shares of choice predictions when alternatives are highly similar, and will assess how well they tolerate the IIA property. Of particular interest is the comparative performance of the RFC rule, which (among other things) was specifically introduced to allow similar alternatives to compete more closely, reducing the IIA bias in preference simulations as a result.

The choice of our experimental factors and the data generation process is based on previous Monte Carlo studies related to conjoint or discrete choice analysis (Andrews, Ainslie & Currim, 2002, Andrews, Ansari, & Currim, 2002; Vriens, Wedel & Wilms, 1996; Wirth, 2010). These previous studies focused on the comparative performance of competing statistical methods for preference estimation, but not on the comparative performance of competing choice rules for share predictions.[3] When verifying whether our findings from the Monte Carlo study also hold under real data conditions, we additionally compare the five choice rules using five empirical data sets. Generally, we expect an improved predictive accuracy for choice rules that use random draws as compared to only point estimates, because the use of draws accounts for preference uncertainty. We further expect an additional benefit from using HB instead of RFC draws since the information about a respondent's choice behavior along with her/his choice task is here implicitly preserved and utilized for choice share predictions.

The article proceeds as follows. We first review previous papers that have dealt with choice rule comparisons in the context of choice-based conjoint analysis or that involve HB draws (Section 2). We then characterize the choice rules in more detail and explain the IIA property (Section 3). This is followed by a description of the experimental design of our Monte Carlo study, specifically introducing the factors that were experimentally varied, explaining the data generation process, and describing the statistical criteria used for comparing true with predicted shares of choice (Section 4). We also present the results of the Monte Carlo study and discuss our findings in Section 4. Finally, we present and discuss the results of applying the five choice rules to empirical data in Section 5. The paper closes with a discussion of our study's implications and limitations, with an outlook for future research (Section 6).

2. Previous research

Early studies comparing different choice rules focused on the traditional first choice and logit choice rules (e.g., Elrod & Kumar, 1989; Finkbeiner, 1988; Green & Krieger, 1988). Even after HB estimation techniques became established (e.g., Allenby & Ginter, 1995; Allenby, Arora & Ginter, 1995; Lenk, Desarbo, Green & Young, 1996) and the RFC choice rule was introduced (Huber et al., 1999), very few studies analyzed the use of HB draws or the RFC rule for choice share predictions. Huber et al. (1999), and Orme and Huber (2000) conducted an empirical CBC study on mid-sized televisions to examine the ability of the RFC choice rule under different levels of variability to correctly reflect similarity effects in market simulations. The authors assessed the impact of the RFC rule applied to the aggregate logit, latent class logit, and individual-level models estimated using the HB mixed logit or Sawtooth Software's ICE (individual choice estimation) method. For each of these models, three levels of variability were examined within the RFC framework: (1) no variability, which is equivalent to using the first choice rule; (2) adding product variability, which is equal to adjusting the scale under the logit choice rule; and (3) adding both product and attribute variability where the latter explicitly corrects for product similarity. Holdout tasks containing extremely similar alternatives were designed to test the performance of the RFC variants to handle different degrees of similarity among products. Predicted combined shares of the near substitute alternatives as a percentage of their actual shares, as well as the mean absolute error (MAE) in predicted choice shares across holdout alternatives were then computed. RFC with product and attribute variability provided the best share predictions, but attribute variability was clearly the crucial component for improving choice share predictions.

Orme and Baker (2000) used the empirical CBC data set collected by Huber et al. (1999) to compare shares of choice predictions based on HB draws versus the RFC choice rule. The mean absolute error (MAE) between actual and predicted choice shares for holdout alternatives was calculated and subsequently scaled as a percentage of the test-retest MAE for repeated choice tasks. The results showed that both using HB draws and applying the RFC rule to individual-level point estimates of part-worths (i.e. adding only attribute variability) worked well. However, predictions based on the RFC rule were slightly more accurate for the data set considered compared to those based on HB draws. Importantly, due to the much slower processing times and limited storage capacities of computers 15 years ago, only 35,000 burn-in iterations were used for HB estimation, which may be too small to ensure convergence of the Markov chain. Similar to Huber et al. (1999), adding attribute variability proved essential for a good performance of the RFC choice rule, while adding product variability in addition to attribute variability did not provide further substantial benefits. It is worth mentioning that Orme (2017) recently reported some new results from the comparison of using HB draws (HBLC) versus the RFC choice rule for shares of choice predictions. Different from the study of Orme and Baker (2000), HB draws worked slightly better than RFC in terms of correlations between predicted choice shares and observed out-of-sample choice shares in holdouts. The findings were based on a CBC data set for vacation cruise packages collected especially for the modeling prize competition launched by Sawtooth Software in 2016 and they were only briefly discussed as a supplement to the main results of this modeling competition. Importantly, no information on the number of burn-in iterations used for HB estimation was provided, and only 200 draws per re-

[3] Vriens et al. (1996) compared different conjoint segmentation methods (two-stage versus integrated approaches); Andrews, Ansari & Currim (2002) compared conjoint models estimated at different levels of aggregation (among others Hierarchical Bayes versus finite mixture models); and Wirth (2010) compared the performance of the HB mixed logit using different levels of information (involving choices of only the best concept versus the best and worst concepts from each choice task).

spondent were used for preference simulation. We further comment on this latter issue in Section 5.

Baier and Polasek (2003) compared the traditional FC rule, the LC rule, and the HBFC rule in the context of metric/ratings-based conjoint data. FC and LC were applied to point estimates of individual part-worths obtained from standard ordinary least squares (OLS) regressions on the one hand and a compositional self-explicated model (SEM) on the other. In addition, individual-level part-worths were estimated using HB regressions, and the generated HB draws were used for first choice simulations (HBFC). Based on holdout validation, the comparative performance of the choice rules was assessed (among others) in terms of Spearman's r between predicted and true choice shares. The results showed a higher predictive accuracy for the HBFC rule compared to the LC rule, albeit a slightly worse predictive performance compared to the traditional FC rule. However, if individual part-worths can be estimated using single OLS regressions (due to sufficient degrees of freedom), the lower level of the HB model should converge against or at least come close to those OLS part-worths. In this special case, it was not unexpected that HB draws didn't outperform first choice predictions based on OLS. And in fact, it is almost impossible to estimate individual-level part-worths based on choice data using separate choice models (MNL models) for each respondent because conjoint choice data provide much less information compared to metric (ratings-based) conjoint data. Holdout validation in the empirical study by Baier and Polasek (2003) was also based on only one single holdout, limiting the generalizability of their results.

Arenoe (2003) conducted ten commercial CBC studies of packaged goods, evaluating the effects of using different estimation and prediction techniques on the external validity of CBC (i.e. the ability of CBC to accurately predict real market shares). The techniques included three methods for utility estimation (aggregate logit, individual choice estimation, and HB mixed logit); three choice rules (FC, RFC with product variability only, RFC with both product and attribute variability); and two correctional measures (weighting respondents by their purchase frequency and weighting estimated product shares by their distribution level). The MAE and the Pearson correlation coefficient between real and predicted shares applied to past market situations were used as measures for external validity. The findings with regard to the choice rules showed that the RFC with product variability error and RFC with product and attribute variability errors outperformed the first choice rule on most occasions. There was also weak evidence that RFC with additional attribute-error correction for similarity outperformed RFC with product variability error only.

Table 1 summarizes the main characteristics and findings of these studies. Only very few studies to date have compared the predictive performance of different choice rules in the context of CBC data. All of them used empirical data, with only one of them investigating the use of HB draws for shares of choice predictions (Orme & Baker, 2000). Furthermore, the findings of this study were based on only one single data set.[4] To the best of our knowledge, no previous study has systematically analyzed the performance of using HB draws in comparison to the RFC rule, the logit choice rule, and the first choice rule for shares of choice predictions. This is why we present in a first step a Monte Carlo study based on synthetic CBC data, which will allow systematic exploration of whether one choice rule is able to provide more accurate share predictions than others under certain market conditions.

One of these conditions is the degree of similarity among products to which the choice share predictions apply. If the degree of similarity between each two alternatives strongly differs (for example, two of three alternatives are highly similar, while the third alternative is not), the IIA property comes into play. Here the question arises regarding how the different choice rules are capable of handling different levels of similarity in the sense that IIA problems can be kept as small as possible.[5] In our Monte Carlo study, we address this question by systematically constructing scenarios where choice sets one time contain two extremely similar, and the other time quite unique (i.e. mutually very dissimilar) alternatives. In the empirical part of this paper, we further compare all five different choice rules based on five different empirical data sets to allow more generalizable implications than previously for real CBC data settings as well. We also verify here whether our findings from the Monte Carlo study hold in empirical settings as well (Section 5). The studies listed in Table 1 can be further distinguished regarding whether they treat the internal or external validity of conjoint simulations. The internal validity is usually accomplished via holdout validation (as in Baier & Polasek, 2003; Huber et al., 1999; Orme & Baker, 2000; Orme & Huber, 2000), whereas the external validity refers to the comparison between predicted shares of choice and real market shares (as in Arenoe, 2003). In our studies, we focus on the internal validity of the different choice rules for preference simulations. For holdout validation we use several measures of predictive accuracy (previous studies only focused on minimizing the MAE; for example, predictive measures that penalize larger prediction errors more strongly (e.g. RMSE) were not considered). The following section describes the five choice rules investigated in this study (FC, LC, RFC, HBFC, HBLC).

3. Choice rules

3.1. First choice rule (FC)

The deterministic FC rule is a very strong choice rule because each respondent is assumed as choosing with certainty the item with the highest predicted utility from a set of alternatives. The FC rule assigns a probability of one to the alternative with the highest utility, and probabilities of zero to all other items, even when utility differences between alternatives are only marginal. As a consequence, shares of choice for alternatives with above-average utility may be strongly overestimated on the aggregate (market) level. The FC rule can however be appropriate for non-routine purchases such as durables like automobiles or personal computers, and high-involvement choice decisions. With frequently purchased consumer goods on the other hand (e.g. beverages), respondents' choice behavior may be more probabilistic, making preferences vary across different use occasions (Elrod & Kumar, 1989; Green & Krieger, 1988; Rao, 2014). The FC rule is furthermore not affected by the IIA property (see below).

3.2. Logit choice rule (LC)

The LC rule assumes that the probability of selecting a product is proportional to the exponential function of its expected utility. It therefore permits continuous probabilities of choice, i.e. each alternative has the chance to be chosen according to its estimated choice probability instead of assigning a probability of 1 to the

[4] Likewise, only one study compared the performance of HB draws for shares of choice predictions to other choice rules based on metric conjoint data (Baier and Polasek 2003, see Table 1).

[5] The IIA property is both an issue for model estimation and choice share predictions. Accommodating preference heterogeneity using the HB mixed logit for utility estimation is known to soften (but not fully eliminate) the IIA property (Train, 2009). Using HB draws appears to provide a natural solution to keep IIA problems small for subsequent shares of choice predictions (see Section 3).

Table 1
Previous studies related to choice rule comparisons based on CBC data or involving HB draws.

Study	Conjoint Approach	Data Basis	Estimated Models	Choice Rules Comparisons	Measures of Predictive Accuracy	Main Findings
Huber, Orme, and Miller (1999) (also see Orme and Huber 2000)	Choice-based	Empirical	Aggregate Logit, Latent Class, HB, ICE	RFC with three levels of variability (no variability, only product variability, both product and attribute variability)	MAE based on holdout tasks, predicted combined shares of near substitutes versus combined actual shares	Adding attribute variability is crucial for a good performance of RFC
Orme and Baker (2000)	Choice-based	Empirical	HB	RFC (attribute variability, both product and attribute variability), HB draws	MAE between actual and predicted shares based on holdout tasks	RFC with attribute variability slightly more accurate than HB draws
Baier and Polasek (2003)	Metric	Empirical	OLS, SEM, HB	FC, LC, HB draws (unrestricted, with ordinal constraints for some attributes)	Spearman's r between actual and predicted shares based on holdout task	Predictive performance of HB draws better than when based on LC, and slightly worse than under FC
Arenoe (2003)	Choice-based	Empirical	Aggregate Logit, HB, ICE	FC, RFC (product variability only, both product and attribute variability)	MAE and Pearson correlation between real and predicted market shares (past market data)	RFC with both product and attribute variability and RFC with only product variability are better than FC
Our approach	Choice-based	Synthetic + Empirical	HB, Aggregate Logit	FC, LC, RFC (with attribute variability), HB draws (HBFC, HBLC)	MAE, MSE, RMSE, MAPE, RAE between true/actual and predicted shares based on holdout tasks	Synthetic data: HB draws applied to choice simulations (HBFC, HBLC) are much more accurate compared to the other choice rules Empirical data: HB draws applied to first choice simulations (HBFC) with the lowest prediction errors

item with the highest utility (Green & Krieger, 1988):

$$P_n(j) = \frac{\exp(\mu \cdot V_{nj})}{\sum_{m=1}^{M} \exp(\mu \cdot V_{nm})}, \quad (1)$$

where $P_n(j)$ is the probability that respondent n chooses alternative j in a particular choice task. V_{nj} is the deterministic utility of alternative j, and $\mu > 0$ represents the scale parameter of the logit model. A weakness of the LC rule is that it suffers from the independence of irrelevant alternatives (IIA) property. This implies that the ratio of probabilities of choosing one product over another remains constant independent of whether further products are available in a choice task or in a competitive market scenario. As a consequence, if a new alternative is introduced, the probabilities for the existing alternatives are reduced by the same percentage, proportional to their pre-introduction preference shares (e.g., Ben-Akiva & Lerman, 1985; Green & Srinivasan, 1978; Jain & Bass, 1989; Louviere & Woodworth, 1983; Louviere, Hensher, Swait & Adamowicz, 2010; Luce, 1959; McFadden, 1974; Ray, 1973; Train, 2009). The IIA property can lead to counterintuitive results if some alternatives are very similar to each other while others are not (Vermeulen, Goos & Vandebroek, 2008). For very similar alternatives, it might be expected that the sum of their predicted shares of choice should be about the same compared to the share of choice if just one of the alternatives were present. However, due to the IIA property, the LC rule overestimates the joint share, resulting in a so called "share inflation" for similar items (Chakraborty, Ball, Gaeth & Jun, 2002; Finkbeiner, 1988). The limitations arising from the IIA property are commonly explained with the classical "red bus-blue bus" paradox (Debreu, 1960). Accordingly, applying the LC rule when larger differences in similarities among alternatives exist can lead to highly distorted forecasts. In this case, cannibalization effects may be underestimated by the predicted joint shares of choice with e.g. a product line extension (e.g., Huber et al., 1999, 2007; Orme & Huber, 2000). In order to combine the advantages and avoid the drawbacks of the FC rule (immune to share inflation, but probably not very realistic) and the LC rule (often more realistic, but prone to the IIA property), the RFC rule as well as the use of HB random draws as an input to the FC rule (HBFC) or to the LC rule (HBLC) present themselves as promising remedies.[6]

3.3. Randomized first choice rule (RFC)

The RFC rule proposed by Huber et al. (1999), and supported by Sawtooth Software modifies the FC rule by adding random error to the point estimates of the part-worths (attribute variability) and/or to each overall product utility (product variability). Adding random components is repeated numerous times, each time introducing new random error terms and subsequently assuming that the respondent chooses the product with the highest utility (Huber et al., 1999, 2007; McCullough, 2002; Orme & Baker, 2000; Orme & Huber, 2000):

$$U_j = (\beta + \varepsilon_\beta)x_j + \varepsilon_j, \quad (2)$$

where U_j is the utility of alternative j, β represents the vector of part-worths, x_j is a dummy vector indicating the presence (= 1) or absence (= 0) of attribute levels of alternative j, ε_β is a random error term added to the part-worths, and ε_j is a random error term added to alternative j. Note that ε_β is held constant across alternatives, while ε_j is unique for each alternative. The random error

[6] IIA problems in preference simulations could also be addressed using a multinomial probit model, recently proposed e.g. by Dotson et al. (2018). However, estimation of probit models is computationally still much more expensive because the resulting integrals have to be evaluated numerically (e.g., see Train, 2009). This becomes even more relevant in Monte Carlo studies with a large number of treatments and replications, as in our study. Note however that the probit model provides a closed-form solution for full conditionals. Therefore, Gibbs sampling instead of the Metropolis-Hastings (MH) algorithm could be used for MCMC estimation, which eliminates the need for tuning MH parameters as required for the logit model (e.g., see Albert & Chib, 1993, Fahrmeir & Kneib, 2011, Chapter 2).

term ε_β introduces correction for product similarity, allowing similar alternatives to compete more closely. If $\varepsilon_j = 0$ then duplicate alternatives always have their share divided in two. If $\varepsilon_\beta = 0$ and ε_j is Gumbel-distributed, then shares of choice predictions after many draws are identical to the LC rule. Since one central focus of this paper is on the capability of the different choice rules to account for similarity relationships between alternatives, and adding product variability clouds these similarity relationships (e.g., Huber et al., 1999), we concentrate on the effects of adding attribute variability only. In doing this, we follow the default setting of Sawtooth Software ($\varepsilon_j = 0$), assuming full correction for product similarity, using their method for auto-calibration of the amount of variance of the error term to disturb the part-worths. It is also important to note that point estimates for respondents resulting from the HB mixed logit model already contain a certain scaling related to the variance of the Gumbel-distributed error term. Adding a new product error term (product variability) here seems unreasonable because it would arbitrarily change the original scaling of the HB utilities. In contrast, adding attribute error to point estimates is reasonable, and reconstitutes some of the variance of the part-worths that was previously eliminated by averaging hundreds or thousands of HB draws to point estimates. As such, shares of choice predictions based on directly using the HB draws of the Markov chain versus RFC draws (i.e. ex-post disturbed point estimates obtained from adding attribute error only) are directly comparable.

3.4. HB random draws as input to the FC or LC rule (HBFC, HBLC)

Using the HB mixed logit model for utility estimation offers the opportunity to directly utilize HB random draws as inputs to FC or LC simulations. In particular, with the use of HB, multiple estimates of part-worths are generated applying Markov Chain Monte Carlo (MCMC) techniques. These estimates are draws from the posterior distribution of part-worths for each respondent. These individual draws can be saved after the so-called burn-in phase (i.e. after convergence of the Markov chain) and used directly as inputs to the FC rule or the LC rule, producing shares of choice predictions for each respondent based on hundreds or thousands of HB draws.[7] Not averaging the individual draws to point estimates preserves the individual scaling of a respondent's part-worths as evolved through the Markov chain (resembling both variances and covariances across part-worths estimated in the upper level of the HB model[8]), as opposed to the RFC rule which adds a new model-independent IID error term to the point estimates *after* having averaged the individual HB draws of a respondent. It is important to ensure via a large number of burn-in iterations that convergence of the Markov chain to the posterior distribution has been achieved (Train, 2009) (see Appendix A). Convergence can be formally tested by using the Gelman-Rubin potential scale reduction factor (PSRF) (see Supplementary Materials). Obviously, when using a proper Bayesian framework like the HB mixed logit model for estimation, any function of parameters should correctly consider the underlying distribution of parameters to derive further quantities of interest (as seen with choice shares in our context). However, as mentioned in the introduction, past and even current research has repeatedly neglected this fact, continuing to use point estimates nevertheless.

Table 2
Advantages and drawbacks of the choice rules.

Choice rule	Advantages	Drawbacks
FC	immune to IIA problems	strict and possibly not very realistic because the product with the highest total utility is assumed to be chosen with certainty
LC	often more realistic due to the chance of choosing any alternative that has a non-zero choice probability	prone to IIA property
RFC	combines benefits of FC and LC, i.e. allows "vote splitting" and correction for product similarity	no theoretical foundation is provided for adding estimation-independent random errors to point estimates
HBFC / HBLC (HB draws+FC, HB draws+LC)	allows "vote splitting" and correction for product similarity, contains more information than RFC draws because the error directly comes from HB estimation, i.e. the individual scaling of a respondent's part-worths is preserved	HB draws must be available (which is e.g. not the case for classical random-effects models or latent class models), saving HB draws requires large storage capacity

Overall, both RFC, HBFC, and HBLC simulate multiple choices per respondent based on underlying distributions of attribute level part-worth parameters.[9] The difference however is that HB draws contain more information because the error directly comes from the HB estimation stage, whereas the RFC method adds a new estimation-independent error after having averaged a respondent's draws from the HB model to point estimates (Huber et al., 2007). So from a statistical point of view, HB provides more empirically correct draws of random error around point estimates. RFC fails to be properly motivated, because no theoretical foundation is provided for adding ex-post estimation-independent random error (Huber et al., 2007). The main advantages and drawbacks of the five choice rules are summarized in Table 2.

4. Design of the Monte Carlo study and results

4.1. Experimental factors

As mentioned above, the choice of our experimental factors and the data generation process derive from previous Monte Carlo studies (Andrews, Ainslie & Currim, 2002, Andrews, Ansari, & Currim, 2002; Vriens et al., 1996; Wirth, 2010). Three experimental factors were manipulated for model estimation, and each factor was varied on several levels, as shown in Table 3.

There were 2^2 x $3 = 12$ experimental conditions for model estimation. With five replications per experimental condition (i.e. six runs per treatment) and subsequent shares of choice predictions based on each of the five choice rules for each preference data set (run), we obtained a total of $12 \times 6 \times 5 = 360$ observations for statistical analysis (i.e. the choice rules were also systematically varied like the other factors). Part-worth utilities were estimated

[7] This means that the FC and LC rules are employed as before, albeit at the draw level (i.e. for each draw of the Markov chain). A respondent's probability of choosing an alternative is obtained by calculating the share of first choices across the HB draws (HBFC), or by averaging the resulting probabilities across the HB draws (HBLC).

[8] https://sawtoothsoftware.com/help/lighthouse-studio/manual/hid_randomizedfirstchoice.html. This means that the RFC rule considers independent draws with equal variance and zero covariances across part-worths.

[9] In random utility models like the HB mixed logit, respondents' choices are assumed to be absolutely deterministic while the researcher might not observe all influencing factors a respondent considers for her/his choice decision. Therefore, from a Bayesian perspective, respondents' repeated choices in conjoint experiments help the researcher to update her/his prior beliefs about respondents' preferences, and these beliefs are represented by the distributions of part-worth parameters used as input for choice share predictions.

Table 3
Experimental factors and factor levels.

Factor	# Factor levels	Factor levels
1. Number of choice tasks (excl. 2 holdouts)	3	7, 11, 15
2. Sample structure	2	less heterogeneous, more heterogeneous
3. Model complexity (number of parameters)	2	simple, complex

using the HB mixed logit model, the standard HB estimation approach for conjoint choice data. The objective was to analyze the performance of the different choice rules, i.e. the first choice (FC) rule, the logit choice (LC) rule, the randomized first choice (RFC) rule with correction for similarity (attribute variability), the first choice rule based on HB draws (HBFC), and the logit choice rule based on HB draws (HBLC) under varying experimental conditions. No previous study has systematically compared these five choice rules. Later, we further add the aggregate MNL model as a benchmark for choice share predictions to assess how much of the bias caused by the IIA property can already be addressed by allowing for unobserved preference heterogeneity across respondents using the HB mixed logit model for estimation (compare Section 4.5).

Based on the results of two meta-analyses, we limited the number of choice tasks per respondent for part-worth estimation to a maximum of 15 (factor 1). Hoogerbrugge and van der Wagt (2006) analyzed 37 HB-CBC studies (i.e. CBC studies where the HB mixed logit was used as an estimation model, as in our study) and observed no substantial improvement in predictive validity when using more than 15 choice tasks. Kurz and Binner (2012) analyzed 12 HB-CBC studies and reported similar findings. In particular, they found that respondents may reach an "Individual Choice Task Threshold (ICT)" beyond which they may become disengaged and tend to use simplification strategies. As a consequence, the predictive performance of HB-CBC may stagnate or even decrease if a larger number of choice tasks per respondent are used. Similar to Hoogerbrugge and van der Wagt (2006), Kurz and Binner recommended using no more than 15 choice tasks.

To compare the predictive performance of the five choice rules in our Monte Carlo study, we included two fixed holdout tasks in addition to 7, 11, and 15 choice tasks used for model estimation. We expected a decrease in predictive accuracy as the number of choice tasks used for model estimation decreased because less information on the individual respondent level would then be available for part-worth estimation. A corresponding significant effect has been confirmed by Wirth (2010) who reported worse shares of choice predictions for a lower number of choice tasks.

Because the HB approach borrows information from the sample population for estimating part-worths on the individual respondent level, it can further be assumed that the amount of preference heterogeneity in the sample affects the individual-level part-worth estimates. Factor 2 considered this aspect, and allows that respondents may be more or less heterogeneous in their part-worth structures (also compare Andrews, Ainslie & Currim, 2002, Andrews, Ansari, & Currim, 2002; Vriens et al., 1996; Wirth, 2010). The findings of the latter three papers[10] indicate that the predictive performance of HB models (significantly) benefits from more heterogeneous samples. Wirth (2010) found that one potential explanation for this unexpected result is the Bayesian shrinkage of individual parameters towards the population mean in HB models, which is ceteris paribus stronger for heterogeneous samples and helps improve shares of choice predictions.

We further manipulated the complexity of the choice task, and differentiate between simple and complex conjoint scenarios (factor 3). The term *simple scenario* means that relatively few parameters need to be estimated (six attributes, three attribute levels), while the *complex scenario* is characterized by a relatively high number of parameters (twelve attributes, five attribute levels). Thus, the total number of parameters to be estimated for each respondent was twelve for the simple condition (two parameters per attribute) and 48 for the complex condition (four parameters per attribute), respectively.[11] We expected the predictive accuracy to be worse for complex compared to simple conjoint scenarios due to the higher number of individual parameters to be estimated. In other words, given a certain amount of information on the individual respondent level (determined by the number of choices per individual), a larger number of parameters would reduce the degrees of freedom for part-worth estimation, and shares of choice predictions would suffer from less reliable part-worth estimates. A significant negative impact of an increasing model complexity on shares of choice predictions was found by Wirth (2010). Since previous Monte Carlo studies reported significant effects of the number of choice tasks, the sample structure, and the model complexity on the predictive performance of HB models, we included them in our study as experimental factors to control for systematic effects that otherwise would be attributed to the choice rules.

Of additional note here is the issue of parameter recovery. Given a fixed number of parameters to be estimated at the individual respondent level, a lower number of choice sets per respondent and therefore less individual information should make it harder to accurately recover the true part-worths. As a result, goodness of parameter recovery may differ depending on experimental conditions, and a worse parameter recovery (expected here for a lower number of choice sets) might negatively affect shares of choice predictions. Similarly, given a certain number of choice tasks per respondent, a higher model complexity is expected to produce a larger bias in parameter recovery. Expressed differently, a larger number of parameters to be estimated given the same amount of individual respondent information should worsen parameter recovery as well as the accuracy of shares of choice predictions. Finally, goodness of parameter recovery could ultimately depend on the sample structure and thus affect the accuracy of shares of choice predictions as well. We later analyze how the accuracy of shares of choice predictions also depends on the goodness of parameter recovery, i.e. the accuracy of recovering the true preferences in the parameter estimation stage under the different experimental conditions (see Appendix D).

Beyond the factors varied in this study, we held other factors common across treatments (data sets) to keep the computational burden manageable. First, we set the sample size to 200 respondents for all treatments. In previous studies, variations in sample size did not show a noticeable effect (if any at all) on the predictive performance of HB models (compare Andrews, Ainslie & Currim, 2002, Andrews, Ansari, & Currim, 2002; Wirth, 2010). Second, we used the standard error variance of the MNL model that corresponds to a scale factor of one. And third, the number of alternatives per choice task was held constant. Both the choice tasks used for model estimation and the two fixed holdout tasks consisted of three alternatives. We further didn't include a no-purchase option

[10] Vriens, Wedel, and Wilms (1996) did not consider HB models.

[11] Note that the part-worth utility for one level of each attribute is set to zero (constituting the reference category).

because considering one provides no additional benefit in the context of simulated data.

4.2. Data generation

The HB mixed logit approach assumes (a) the MNL model as a discrete choice model for each respondent (lower level model) and (b) a multivariate normal distribution $\beta_n \sim N(\bar{\beta}, V_\beta)$ for the population of respondents (upper level model).[12] Following previous Monte Carlo studies, V_β was specified as a diagonal matrix (compare e.g. Andrews, Ainslie & Currim, 2002, Andrews, Ansari, & Currim, 2002; Wirth, 2010). The data generation process was based on Wirth (2010), and followed Wirth's experiences about typical distributions of average part-worths in real-world applications. Accordingly, for 80% of the attribute levels the corresponding population mean betas (contained in $\bar{\beta}$) were randomly drawn from the interval [−2, 2], and for each 10% of the attribute levels randomly from the intervals [−5, −2] and [2, 5]. The covariance matrix V_β captures the amount of preference heterogeneity across respondents and was generated from a combination of gamma and uniform draws, approximating a multivariate normal preference distribution as a result. Depending on the specification of the gamma and uniform distributions, more or less heterogeneous samples were obtained (factor 2). In addition, the variances along the main diagonal of V_β were allowed to differ between attribute levels (for details on generating the covariance matrix see Wirth (2010)). The resulting distributions of individual-level part-worths for a less heterogeneous and for a more heterogeneous sample are exemplarily illustrated in Appendix B, Fig. B.1. As can be seen from the density plot in Fig. B.1, part-worth utilities for each attribute level are normally distributed across respondents with different variances across attribute levels, thus representing a multivariate normal distribution of preference heterogeneity (as specified for the data generation process).

After the "true" individual-level part-worths β_n were generated, deterministic utilities of respondent n for each alternative in each choice task were obtained from computing the linear combination $U_n = X\beta_n$. Choice tasks were generated using the complete enumeration algorithm implemented in the Sawtooth software. Lastly, a Gumbel-distributed error term was added to U_n to obtain the stochastic utilities (using the standard error variance of the MNL). Based on the simulated choices for 200 respondents, the individual-level part-worths (as well as the population mean betas $\bar{\beta}$ and the covariance matrix V_β) were then re-estimated using the HB mixed logit model.

The two holdout tasks with three alternatives (products) were carefully designed for each treatment. To address the potential IIA bias, one holdout task was designed to have two of the three alternatives extremely similar to each other in their profiles. In contrast, the second holdout task consisted of alternatives that were all different from each other. The holdout tasks were generated as follows: First, we designed a holdout task with two alternatives A and B such that the shares of choice of the two alternatives were about 70% for A and 30% for B. Next, a third alternative C was introduced that was very similar to alternative A. Similarity was achieved by modifying one attribute level of alternative A at a time. The sum of deterministic utilities across respondents was computed for each possible combination of attribute level modifications. Finally, we decided upon the attribute level combination for alternative C that minimized the utility difference between alternative A and C across respondents.[13] Alternatively, we designed a holdout task with a third alternative C' where the combination of attribute levels was as different as possible from both alternatives A and B. Finally, we separately compared the true to the predicted shares of choice for each of the two holdout tasks, resulting in 360 observations for each holdout scenario.

4.3. Measures of performance

No single best forecasting measure exists according to Winkler and Murphy (1992). It is therefore reasonable to use alternative statistics for measuring predictive accuracy, each of them having somewhat different strengths and weaknesses. Under each treatment, we compared the predicted shares of choice ($\hat{W}_j$) based on the re-estimated part-worths to the "true" shares of choice (W_j) based on the generated part-worths across alternatives j in holdout task k along the following five measures of predictive accuracy (Leeflang, Wittink, Wedel & Naert, 2000):

The *Mean Absolute Error (MAE)* measures the average absolute deviation between true and predicted shares of choice:

$$\mathrm{MAE\,(W)} = \frac{1}{J}\sum_j \left|\hat{W}_j - W_j\right| \tag{3}$$

By squaring the deviations, the *Mean Squared Error (MSE)* weights large prediction errors more heavily than small prediction errors. The disproportionate influence of larger deviations is a desirable property, because larger prediction errors with regard to shares have disproportionate negative effects on managerial decisions (Chakraborty et al., 2002):

$$\mathrm{MSE(W)} = \frac{1}{J}\sum_j \left(\hat{W}_j - W_j\right)^2 \tag{4}$$

Taking the square root of the MSE yields the *Root Mean Squared Error (RMSE)*. Like the MSE, the RMSE penalizes larger prediction errors more strongly but finally reports the average prediction error in the dimension of the original measurement units. Therefore, the RMSE is directly interpretable in terms of measurement units:

$$\mathrm{RMSE(W)} = \sqrt{\frac{\sum_j \left(\hat{W}_j - W_j\right)^2}{J}} \tag{5}$$

The *Mean Absolute Percentage Error (MAPE)* is a dimensionless measure. Similar to the MAE, absolute rather than squared prediction errors are computed. However, each absolute prediction error is expressed relative to the true choice share for alternative *j*:

$$\mathrm{MAPE\,(W)} = \frac{1}{J}\sum_j \left|\frac{\hat{W}_j - W_j}{W_j}\right| \cdot 100\% \tag{6}$$

The *Relative Absolute Error (RAE)* compares the prediction error of a given model to the prediction error obtained from a naive forecasting model, where the latter defines a benchmark by using choice probabilities that would result purely by chance (as denoted by B_j)[14]:

$$\mathrm{RAE(W)} = \frac{\sum_j \left|\hat{W}_j - W_j\right|}{\sum_j \left|B_j - W_j\right|} \tag{7}$$

If the RAE statistic is less than one the forecasting model outperforms the chance model, and if the RAE statistic is greater than one the forecasting model is worse than the chance model.

[12] The HB mixed logit is described in greater detail in Appendix A.

[13] Note that the construction of the holdout tasks was based on the generated "true" individual-level part-worths. Summing up deterministic utilities across respondents would not be allowed based on (re-)estimated individual-level part-worths due to the individual scaling of the parameter estimates.

[14] Having three alternatives in each holdout task, the choice probability due to chance was 33.33% for each alternative.

Values of zero indicate no prediction error for all five measures of predictive accuracy, i.e. a perfect prediction of the "true" choice shares.

4.4. Discussion of results

The impact of the number of choice tasks, the sample structure, the model complexity, and in particular the type of choice rule on each of the five dependent measures for the accuracy of shares of choice predictions was investigated by analyses of variance for both main effects and first-order interaction effects (between factors). Recalling that the number of choice tasks, the sample structure, and the model complexity were manipulated for the model estimation stage, and the performance of the different choice rules on shares of choice predictions were potentially moderated by the generated condition, conducting analyses of variance enables us to investigate the main effect of the type of choice rule (i.e. averaged over the manipulated conditions for parameter estimation); the main effects of each of the three factors choice tasks, sample structure and model complexity; as well as first-order interaction effects between the type of choice rule and the number of choice tasks, sample structure, or model complexity (i.e. whether differences in the predictive performance of the choice rules critically depend on the manipulated conditions) on the accuracy of shares of choice predictions.

There were two different holdout choice scenarios, as noted above. One was designed to have two of the three alternatives extremely similar to each other in utility (SIMILAR). The second holdout choice scenario consisted of three alternatives that were all different from each other (DISSIMILAR). A total of ten ANOVAs (five measures of predictive accuracy times two holdout scenarios) were conducted accordingly, each based on 360 observations with 330 degrees of freedom for error (within-group degrees of freedom). A first inspection of the ANOVA results revealed that the variance homogeneity assumption was often not met. It is well known that the F-test is fairly robust against violations of the assumption of variance homogeneity when group sizes are equal, as is the case in our setting with six observations for each group (Box, 1954; Glass, Peckham & Sanders, 1972). Nevertheless, we applied the Box correction for heterogeneous variances (Box, 1954) to ensure that this violation did not seriously affect the F-statistic. The Box approximation corrects the F-statistic by adjusting the degrees of freedom (see Supplementary Materials).

F-tests for main and interaction effects and corresponding p-values as well as the explained variances from the ANOVAs for each performance measure are provided as Supplementary materials (see Tables S1 and S2). As seen, the F-test was highly robust because there were no substantial differences with regard to the significance of main and first-order interactions effects with versus without the Box correction. This allowed us to proceed to the next step of summarizing the ANOVA results. All main effects turned out highly significant, independent of whether two very similar alternatives were contained in the holdout choice scenario or not. With regard to first-order interactions, we observed significant effects between the factors sample structure and model complexity (sample structure x model complexity) and between the factors sample structure and number of choice tasks (sample structure x number of choice tasks) for both types of holdout choice scenarios and across nearly all predictive measures.

Both interaction effects are depicted in Fig. 1 exemplarily for the holdout choice scenarios of the DISSIMILAR type and the MAE measure. Considering the *sample structure x model complexity* interaction on the left-hand side of Fig. 1, the patterns of means show that given a complex CBC model (twelve attributes, five levels) the sample structure of respondents (more heterogeneous versus less heterogeneous) had virtually no effect on predictive accuracy: the MAE was stable under the two conditions (as shown by the near-horizontal line). However, for a less complex CBC model, predictive accuracy can be strongly improved when the sample structure is more heterogeneous. Stated differently, a model with few parameters (part-worths) combined with a higher level of respondent heterogeneity has a large positive effect on the predictive accuracy. Therefore, a higher level of heterogeneity across respondents is beneficial for choice share predictions.[15] The interaction *sample structure x number of choice tasks* is illustrated on the right-hand side of Fig. 1. In general, and as expected, a decrease in the number of choice tasks led to worse share predictions, (also see the discussion on main effects below). However, for a less heterogeneous respondent sample, decreasing the number of choice tasks had a much stronger negative effect on predictive accuracy compared to a more heterogeneous sample. Again, a higher level of heterogeneity across respondents seems beneficial for shares of choice predictions. Nearly the same patterns of means could be observed for holdout choice scenarios of the SIMILAR type. Findings of previous Monte Carlo studies have so far only suggested that the predictive performance of HB models might benefit from more heterogeneous samples (Andrews, Ainslie & Currim, 2002, Andrews, Ansari, & Currim, 2002; Wirth, 2010), but have not provided any differentiated views concerning these two interactions. In contrast, 90% of all interaction effects between the type of choice rule and the other factors (across both types of holdout choice scenarios) were not significant, suggesting that differences in the performance of the choice rules do not critically depend on the experimental condition (compare Supplementary Materials, Tables S1 and S2).

Table 4
Effect sizes of main and interaction effects for holdout choice scenarios where two out of three alternatives are extremely SIMILAR measured by eta squared (η^2).

	Predictive Accuracy				
Source (d.f.)	MAE	MSE	RMSE	MAPE	RAE
Choice rule (4)	.229	.171	.226	.248	.228
Number of choice tasks (2)	.112	.094	.117	.124	.112
Sample structure (1)	.013	.013	.013	.023	.013
Model complexity (1)	.117	.102	.121	.101	.117
Sample structure x Model complexity	.014	.004	.016	.014	.014
Sample structure x Number of choice tasks	.009	.013	.008	.012	.009

We calculated eta squared (η^2) as a measure of effect size in ANOVA to further assess the relevance of the main and first-order interaction effects. The corresponding effect sizes are reported in Tables 4 and 5. We followed Cohen's (1988) guidelines to interpret the effect size ($\eta^2 = 0.01$ small effect, $\eta^2 = 0.06$ medium effect, $\eta^2 = 0.14$ large effect). For the holdout choice scenarios with two extremely similar alternatives (Table 4, SIMILAR), we observe very large effect sizes for the choice rule with respect to all performance measures, as well as medium effect sizes for the number of choice tasks and the model complexity. For holdout choice scenarios of the DISSIMILAR type (Table 5), the effect sizes of the choice rule are still larger, except for the MSE. In contrast, the effect sizes of the number of choice tasks and the model complexity, although still medium, are somewhat smaller. The sample structure as well as the two first-order interactions discussed above show smaller

[15] The plots for the *sample structure x model complexity* interactions are identical or look at least very similar for all other predictive accuracy measures.

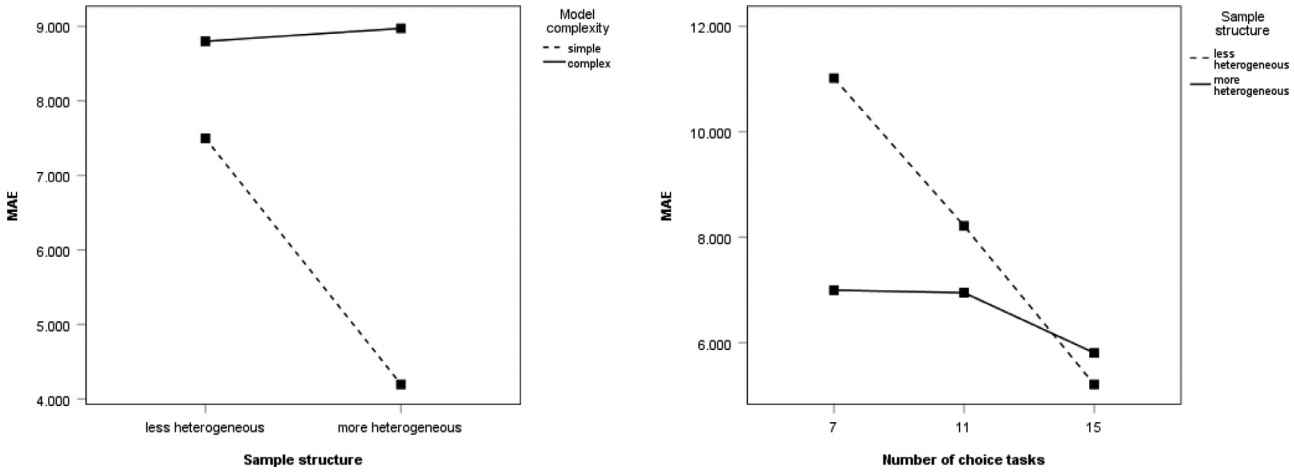

Fig. 1. Interaction effects between the factors *sample structure* and *model complexity* (left-hand side) and between the factors *number of choice tasks* and *sample structure* (right-hand side) for holdout choice scenarios with DISSIMILAR alternatives (exemplarily for the MAE statistic).

Table 5
Effect sizes of main and interaction effects for holdout choice scenarios with DISSIMILAR alternatives measured by eta squared (η^2).

	Predictive Accuracy				
Source (d.f.)	MAE	MSE	RMSE	MAPE	RAE
Choice rule (4)	.248	.170	.251	.322	.248
Number of choice tasks (2)	.078	.074	.078	.078	.078
Sample structure (1)	.023	.030	.024	.065	.023
Model complexity (1)	.088	.069	.092	.080	.088
Sample structure x Model complexity	.029	.013	.026	.004	.029
Sample structure x Number of choice tasks	.034	.027	.035	.013	.034

effect sizes under both types of holdout choice scenarios.[16] Overall, the type of choice rule, and the number of choice tasks and the model complexity appear to be primary drivers for shares of choice predictions.

Tables 6 and 7 display the means of the five measures of predictive accuracy depending on the experimental condition and type of choice rule. For factors with more than two levels, post hoc tests were conducted to examine which of the factor level means significantly differed from each other. The Bonferroni correction was used for the post hoc tests.

The following first focuses on holdout choice scenarios of the SIMILAR type that should be much more prone to an IIA bias due to the fact that they contain two extremely similar alternatives. Subsequently, we compare these results to those obtained for holdout choice scenarios of the type DISSIMILAR where the IIA property should be of less concern, as here all three alternatives are quite different from each other. In general, the results in Table 6 reveal a significant impact of all factors on all measures of predictive accuracy. As expected, the performance measures become considerably worse when the CBC model is complex (twelve attributes, five attribute levels) rather than simple (six attributes, three attribute levels). Further, prediction errors significantly increase with smaller numbers of choice tasks. The smaller the number of choice tasks per respondent given the number of parameters to be estimated, the more information across respondents is pooled, i.e. the more the individual-level part-worths are shrinked towards the population mean. As a consequence, because of less heterogeneity in estimated part-worths, the IIA property is more prevalent because less heterogeneity coincides with more IIA problems (compare e.g. Orme, 1998). Importantly, the decrease in predictive accuracy is much more pronounced when the number of choice tasks is reduced from 15 to 11 compared to a further reduction from 11 to 7. In addition, with respect to the sample structure, the results show that a more heterogeneous sample structure leads to significantly lower prediction errors than a less heterogeneous one (but note that the factor sample structure showed only small effect sizes on the measures of predictive accuracy).

With regard to the choice rule, the most important finding is that both the HBFC rule and the HBLC rule, i.e. using HB draws as inputs to the FC rule or LC rule, by far provides the lowest prediction errors. Both choice rules lead to comparably low prediction errors that do not differ significantly across all five performance measures, with HBLC performing slightly better.[17] On the one hand, the use of HB draws repeatedly applied to first choice simulations appears to soften the strictness of the FC rule considerably. On the other hand, the use of HB draws repeatedly applied to logit choice simulations seems to strongly soften the IIA property. In comparison to using HB draws, the accuracy of shares of choice predictions is considerably worse under the FC, LC, and RFC rule, respectively. As mentioned, the LC rule suffers from the IIA property to a degree that the total share of the two similar alternatives is overestimated, leading to an inflated joint share for the two alternatives. Although not suffering from the IIA property, share predictions based on the FC rule are still worse than based on the LC rule, and the FC rule leads to the highest prediction er-

[16] As mentioned above, 90% (or 27 out of 30) of the interaction effects between the type of choice rule and any of the three factors (number of choice tasks, sample structure, and model complexity) were not significant, and the remaining three interactions showed only small effect sizes between 0.021 and 0.028. In addition, two of these were significant according to the F-test ($p \leq .05$), but not the Box correction ($p \geq .05$). At the same time, these were the only two effects where the F-test and Box approximation diverge in their statistical implications (see Supplementary Materials, Tables S1 and S2). This again emphasizes how differences in the performance of the choice rules do not critically depend on the manipulated conditions.

[17] As mentioned in footnote 7, a respondent's probability of choosing an alternative using the HBLC rule was obtained by averaging over the resulting probabilities across the HB draws. Alternatively, a multinomial draw could be taken for each of the draws based on the draw-specific probabilities for the alternatives to decide which alternative would be chosen each time. Prediction errors for shares of choice predictions turned out identical up to the first and often even the second decimal place for nearly all means of performance measures in Tables 6 and 7. We would like to thank an anonymous reviewer for pointing out the multinomial draw variant.

Table 6
Means of performance measures by experimental condition (i.e. for each factor level)[a] for holdout choice scenarios where two out of three alternatives are extremely SIMILAR.

	Predictive Accuracy				
Factor	MAE	MSE	RMSE	MAPE	RAE
Choice rule					
(1) FC	$11.134^{4,5**}$	$190.009^{4,5**}$	$12.274^{4,5**}$	$36.048^{4,5**}$	$.337^{4,5**}$
(2) LC	$9.129^{4,5**}$	$131.000^{4,5**}$	$10.187^{4,5**}$	$30.414^{4,5**}$	$.277^{4,5**}$
(3) RFC	$9.759^{4,5**}$	$150.631^{4,5**}$	$10.802^{4,5**}$	$31.829^{4,5**}$	$.296^{4,5**}$
(4) HBFC	$5.212^{1,2,3**}$	$49.447^{1,2,3**}$	$5.860^{1,2,3**}$	$16.395^{1,2,3**}$	$.158^{1,2,3**}$
(5) HBLC	$4.825^{1,2,3**}$	$43.487^{1,2,3**}$	$5.425^{1,2,3**}$	$15.304^{1,2,3**}$	$.147^{1,2,3**}$
Number of choice tasks					
7	9.677^{15**}	156.998^{15**}	10.779^{15**}	31.576^{15**}	$.293^{15**}$
11	8.798^{15**}	126.674^{15**}	9.785^{15**}	28.761^{15**}	$.267^{15**}$
15	$5.560^{7,11**}$	$55.072^{7,11**}$	$6.164^{7,11**}$	$17.657^{7,11**}$	$.169^{7,11**}$
Sample structure					
Homogeneous	8.616^{*}	128.649^{*}	9.572^{*}	28.576^{**}	$.261^{*}$
Heterogeneous	7.408^{*}	97.181^{*}	8.247^{*}	23.420^{**}	$.224^{*}$
Model complexity					
Simple	6.202^{**}	68.427^{**}	6.898^{**}	20.573^{**}	$.188^{**}$
Complex	9.822^{**}	157.403^{**}	10.921^{**}	31.423^{**}	$.298^{**}$

*Indicates that the difference between two means is significant at the .05 level (as indicated by the Bonferroni correction method).
**Indicates that the difference between two means is significant at the .01 level (as indicated by the Bonferroni correction method).
[a]Superscripts on means refer to the factor levels and indicate the factor level that leads to a significant difference between the means of the measure of performance.

Table 7
Means of performance measures by experimental condition (i.e. for each factor level)[a] for holdout choice scenarios with DISSIMILAR alternatives.

	Predictive Accuracy				
Factor	MAE	MSE	RMSE	MAPE	RAE
Choice rule					
(1) FC	$9.986^{4,5**}$	$154.685^{4,5**}$	$11.096^{4,5**}$	$46.370^{4,5**}$	$.303^{4,5**}$
(2) LC	$9.067^{4,5**}$	$132.377^{4,5**}$	$10.077^{4,5**}$	$42.889^{4,5**}$	$.275^{4,5**}$
(3) RFC	$9.253^{4,5**}$	$137.475^{4,5**}$	$10.290^{4,5**}$	$43.476^{4,5**}$	$.280^{4,5**}$
(4) HBFC	$4.386^{1,2,3*}$	$37.117^{1,2,3**}$	$4.874^{1,2,3**}$	$22.472^{1,2,3**}$	$.133^{1,2,3**}$
(5) HBLC	$4.136^{1,2,3*}$	$33.698^{1,2,3**}$	$4.605^{1,2,3**}$	$21.688^{1,2,3**}$	$.126^{1,2,3**}$
Number of choice tasks					
7	9.006^{15**}	137.426^{15**}	9.949^{15**}	39.941^{15**}	$.273^{15**}$
11	7.581^{15**}	106.304^{15**}	8.496^{15**}	38.387^{15**}	$.230^{15**}$
15	$5.509^{7,11**}$	$53.480^{7,11**}$	$6.120^{7,11**}$	$27.809^{7,11**}$	$.167^{7,11**}$
Sample structure					
Homogeneous	8.147^{**}	121.328^{**}	9.066^{**}	40.308^{**}	$.247^{**}$
Heterogeneous	6.584^{**}	76.813^{**}	7.311^{**}	30.450^{**}	$.199^{**}$
Model complexity					
Simple	5.846^{**}	65.639^{**}	6.476^{**}	29.943^{**}	$.177^{**}$
Complex	8.885^{**}	132.501^{**}	9.901^{**}	40.815^{**}	$.269^{**}$

*Indicates that the difference between two means is significant at the .05 level (as indicated by the Bonferroni correction method).
**Indicates that the difference between two means is significant at the .01 level (as indicated by the Bonferroni correction method).
[a]Superscripts on means refer to the factor levels and indicate the factor level that leads to a significant difference between the means of the measure of performance.

rors independent of which predictive validity measure is used. The accuracy of choice share predictions based on the RFC rule is not significantly different from that based on the FC and the LC rules. Consequently, in contrast to directly using HB draws for first choice simulations, generating draws by adding estimation-independent random errors to point estimates does not seem to improve shares of choice predictions compared to the simpler LC and FC rules. We did not expect that worse performance by the RFC rule compared to the HBFC and HBLC rules, particularly given the empirical findings by Orme and Baker (2000) (see Section 2. Importantly, these findings were based on one single empirical data set).

Similar findings were obtained for the holdout choice scenarios of the DISSIMILAR type (Table 7). Here, the ANOVA results also indicate a statistically significant impact of all factors on all measures of predictive accuracy. However, compared to the holdout choice scenarios of the type SIMILAR, we observe that the means of nearly all predictive measures are somewhat lower. Interestingly, the means of the MAPE statistic are consistently higher than in Table 6. Since the MAPE measures prediction errors relative to the true choice share, they are more heavily weighted when true choice shares are small. In fact, the true choice share of the third alternative C usually turned out lower when it was constructed as pairwise different from the alternatives A and B (DISSIMILAR scenarios) compared to the case where it was constructed to be highly similar to alternative A (SIMILAR scenarios). This occurred because alternative A was the item with the higher share among alternatives A and B, thus alternative C as a rule obtained a higher true share in scenarios of the type SIMILAR when it was designed to be highly similar to alternative A. Consequently, rather small true preference shares occurred more often in holdout choice scenarios

of the type DISSIMILAR, leading to higher means of the MAPE. RAE values are low for both types of holdout choice scenarios (SIMILAR, DISSIMILAR) indicating that all choice rules provide much better shares of choice predictions than the chance model. The RAE is a bit better (lower) when the alternatives are all different from each other.

Further, we observe noticeable differences compared to holdout choice scenarios of the type SIMILAR with respect to the performance of the choice rules. First, the FC rule performs much closer to the LC and RFC rules. Since the three alternatives were specified as very different in their profiles, and hence utility differences between alternatives usually turned out to be not too small, the FC rule is more appropriate for predicting choices. Moreover, the results show that prediction errors under the LC rule are only slightly lower compared to holdout choice scenarios with two similar alternatives (except for the MAPE). This finding indicates that accounting for heterogeneity by using the HB mixed logit model for utility estimation in general softens the IIA property, because otherwise prediction errors under the LC rule should have turned out higher for holdout choice scenarios of the SIMILAR type.

At this point, the most important result of our study is that using a choice rule based on HB draws (HBFC or HBLC) clearly provides the lowest prediction errors compared to the other choice rules investigated. This finding holds, independent of whether holdout choice scenarios contain two extremely similar alternatives (making predictions prone to an IIA bias) or not. We did in fact expect the superiority of using HB draws for shares of choice predictions based on the provided arguments (see Section 3), just not with this level of clarity. We did not expect that the RFC rule would perform considerably worse compared to the HBFC and HBLC rules, nor that the RFC rule would perform even worse than the LC rule, especially when two of the three holdout alternatives were extremely similar (i.e. when they provide nearly the same utility to respondents). As a consequence, these findings should encourage practitioners to directly use HB draws for preference simulations.

4.5. Refinement

We extended our Monte Carlo study by additionally including the aggregate MNL model to further assess how much the consideration of preference heterogeneity per se contributes to improve shares of choice predictions (the aggregate MNL was excluded in the first step because it completely ignores respondent heterogeneity). Consequently, we expected a huge error for predictions based on the aggregate MNL, which should considerably affect the ANOVA results, including with regard to the other factors and their levels. In addition, the aggregate MNL is fully prone to the IIA property.

Accordingly, we generated data for $2^2 \times 3 = 12$ additional treatments (compare Table 3), again with five replications per experimental condition (a total of six runs), obtaining an additional 72 observations for statistical analysis. For the aggregate MNL, part-worths were determined via traditional maximum likelihood estimation, with shares of choice consistently computed using the logit choice (LC) rule. The means of our five performance measures, depending on the type of holdout choice scenario, the number of choice tasks, the sample structure, the model complexity, and the type of choice rule used, and now including predictions from the aggregate MNL model, are summarized in Appendix C, Tables C.1 and C.2. Both tables show that shares of choice predictions based on the aggregate MNL model (values indicated in bold) turn out extremely badly. As a rule, prediction errors under the aggregate MNL model are at least twice as high compared to any other of the previously considered choice rules that were all applied to individual-level part-worths. Generally, prediction errors for the factors number of choice tasks, sample structure, and model complexity increase under each factor level simply because the much higher prediction error of the aggregate MNL is additionally distributed across factors and factor levels.

It is well known that accounting for heterogeneity can further reduce IIA problems, so that when choices are aggregated across individual respondents, aggregate shares do not strictly adhere to the IIA property (see Alvarez & Nagler, 1998; Ben-Akiva & Lerman, 1985; McFadden, Train & Tye, 1977; Orme, 1998). As seen in Appendix C, Tables C.1, and C.2, the prediction errors under the aggregate MNL (and using the LC rule) are much higher for holdout choice scenarios of the type SIMILAR (except for the MAPE), as we expected due to the larger influence of the IIA property for holdout scenarios of that type. However, having accounted for heterogeneity by estimation of individual-level part-worths using the HB mixed logit, prediction errors occurring as a result of using the LC rule largely decreased to comparably low error levels in both holdout choice scenarios (e.g. for the MAE from 23.258 with the aggregate MNL to 9.129 for the HB mixed logit in holdout choice scenarios with similar alternatives, see Table C.1). This indicates that accommodating individual heterogeneity already remedies the IIA bias to a greater extent, even if point estimates are used from the HB mixed logit estimation as done with the LC rule.[18] However, using the HBLC rule (but also HBFC) instead of the logit choice rule (LC) provides an additional boost in the accuracy of choice share predictions by balancing out much more of the remaining IIA bias (e.g. the MAE considerably decreases further from 9.129 for the LC rule to 4.825 for the HBLC rule, see Table C.1). Stated differently, prediction errors can be largely reduced even further when directly using the draws from the posterior distribution of the estimated HB model, which is in line with a fully Bayesian approach. Note that some IIA bias is also prevalent in holdout choice scenarios of the type DISSIMILAR, otherwise the LC rule should have provided still lower prediction errors here. The reason for this additional improvement in predictive accuracy through the use of HB draws is that model-inherent uncertainty in the choice behavior of the respondents can be taken into account for preference simulation, further reducing IIA problems. Recalling that the RFC rule was introduced among other things to allow similar alternatives to compete more closely and reduce the IIA bias as a result. The power of using HB draws to reduce the IIA bias is nevertheless much higher according to our results from the Monte Carlo study. Finally, Tables C.1 and C.2 demonstrate that the much better performance when using HB draws for shares of choice predictions compared to the other choice rules is highly robust against differences in similarities between options. There is obviously no alternative to using the posterior distribution of estimated parameters to minimize prediction errors. In other words, the HB mixed logit approach deals very well with shares of choice predictions for both dissimilar and similar alternatives if it is consistently applied in a Bayesian sense by directly using the individual HB draws from the estimated model for preference simulations (HBFC, HBLC).[19]

5. Empirical study

Findings from Monte Carlo studies with artificial data can differ from those of empirical studies with real data. If artificial data contain inadvertent biases and do not adequately reproduce the choice behavior of real respondents, findings from Monte Carlo

[18] Dotson et al. (2018) recently demonstrated that accounting for heterogeneity does not fully correct for the IIA in the HB mixed logit, hence "heterogeneity does not "fix" the problem of IIA, as often assumed in marketing and applied economics literature" (Dotson et al. 2018, p. 43).

[19] We provide details on computation times for parameter estimation and preference simulation in Appendix E.

Table 8
Data set characteristics.

Characteristics	Data sets				
	Drill hammer	Glass insurance	Summer tires	Lottery	Tablet
Country	Spain	Netherlands	Germany	Belgium	United Kingdom
Year	2009	2010	2011	2009	2010
Respondents	280	496	662	413	1000
Choice tasks	15	15	8	15	13
Alternatives	4	4	4	4	4
Attributes	8	6	17	8	14
Parameters	12	16	94	30	43

studies can be misleading and not generalizable to empirical settings, and bear the risk of providing wrong implications in practical applications.[20] For example, real respondents may use simplification strategies by concentrating on key attributes like brand or price, undermining the assumption of compensatory choice behavior assumed in all previous Monte Carlo studies. On the other hand, some respondents might become bored during later choice tasks, leading to a higher degree of stochastic choice behavior than considered with artificial data (e.g., Selka, Baier & Kurz, 2014).

To test if our findings from the Monte Carlo study (which clearly favor the use of HB draws) also hold in real-life scenarios, we compared the performance of the choice rules for shares of choice predictions based on five empirical data sets. The empirical data sets were provided by Kantar TNS, one of the world's largest market research institutes. The characteristics of the data sets with regard to the CBC design used (number of choice tasks, number of alternatives per choice task, number of attributes, number of parameters, number of respondents), the product category, year of data collection, and country are summarized in Table 8.

The data sets vary with respect to the number of choice tasks (8 to 15), number of attributes (6 to 17), and number of parameters (12 to 94). Note that we included these three design parameters with comparable ranges in our Monte Carlo study as experimental factors (compare Table 3). The only exception is the summer tires study, which involves a higher number of attributes and a much larger number of parameters to be estimated. In addition, the five CBC studies cover very different product categories and were conducted in five different countries. In all data sets, the choice tasks were composed of three "real" alternatives and a no-purchase option.

The drill hammer study was conducted to determine whether and which value added services can justify a markup. Optimal new product design to attract as many customers as possible was the object of investigation both in the glass insurance and the lottery studies. The starting point for the summer tires study was the EU-wide introduction of different labels for summer tire classification. The objective of the study was to explore if and how the new EU classifications affect customers' perceptions of summer tire offerings. Finally, the tablet study was set up to analyze which computer components may compensate for a loss in utility resulting from not being able to offer well-known operating systems such as iOS or Android. All five empirical studies shared similar objectives, i.e. finding new product designs or modifying/extending existing product designs to eventually increase market share, or revenues, or contribution to profit from sales to a population of target customers. Note that accurately predicting shares is as well essential for revenue or contribution to profit calculations because preference/market shares are integral part of them.

We again performed holdout validation to compare the capabilities of the different choice rules for shares of choice predictions in empirical settings. Here we repeatedly (five times in each data set) chose one random choice task to represent a holdout, ran the HB mixed logit model based on the remaining choice tasks, and calculated our five performance measures (MAE, MSE, RMSE, MAPE, RAE) for the holdout alternatives.[21] We checked the convergence of the Markov chains to the posterior distributions via the potential scale reduction factor (Brooks & Gelman, 1998), with corresponding PSRF values of $R < 1.1$ for the drill hammer, glass insurance, and lottery data, and PSRF values of $R < 1.2$ for the summer tires and tablet data.

Table 9 displays the results of the holdout validation aggregated across the five random holdout choice tasks per data set. It provides a clear picture about the performance of the different choice rules, confirming the main finding of our Monte Carlo study that the use of HB draws should be the method of choice for shares

Table 9
Predictive performance of choice rules (shares of choice predictions) in empirical settings.

		Predictive Accuracy[a]				
Data set	**Choice rule**	MAE	MSE	RMSE	MAPE	RAE
Drill hammer						
	(1) FC	3.49	21.14	4.14	17.47	.34
	(2) LC	2.85	12.25	3.35	18.98	.28
	(3) RFC	2.66	10.74	3.09	15.40	.26
	(4) HBFC	**2.47**	**10.12**	**3.02**	**11.61**	**.24**
	(5) HBLC	2.59	10.86	3.05	16.86	.25
Glass insurance						
	(1) FC	2.81	15.54	3.43	11.73	1.02
	(2) LC	2.98	17.25	3.58	12.61	1.02
	(3) RFC	2.58	13.06	3.06	11.17	.88
	(4) HBFC	**2.07**	**10.81**	**2.49**	**9.05**	**.72**
	(5) HBLC	2.51	13.56	3.07	10.75	.86
Summer tires						
	(1) FC	2.82	13.32	3.35	12.23	1.02
	(2) LC	2.76	13.42	3.32	12.05	.94
	(3) RFC	1.64	4.39	1.99	6.99	.77
	(4) HBFC	**1.08**	**2.34**	**1.35**	**4.64**	**.50**
	(5) HBLC	1.88	6.55	2.28	8.16	.66
Lottery						
	(1) FC	2.13	7.14	2.53	8.74	.43
	(2) LC	1.57	3.76	1.85	6.44	.31
	(3) RFC	1.46	3.45	1.70	5.84	.29
	(4) HBFC	**1.43**	**3.33**	**1.65**	**5.74**	**.28**
	(5) HBLC	1.67	4.35	1.93	7.04	.32
Tablets						
	(1) FC	4.29	30.75	5.13	17.85	1.64
	(2) LC	3.95	28.69	4.80	16.60	1.48
	(3) RFC	2.86	16.05	3.48	12.16	1.01
	(4) HBFC	**2.81**	**15.81**	**3.41**	**11.95**	**.98**
	(5) HBLC	3.36	22.81	4.14	14.25	1.24

[a]Bold values indicate the best-performing choice rule. Underlines indicate that the performance of a choice rule is not statistically different from the best-performing rule ($p > .05$).

[20] We would like to thank two anonymous referees for pointing this out.

[21] We would like to thank one of the anonymous referee for suggesting this kind of holdout validation.

of choice predictions in preference simulations. In all five empirical settings, the HBFC rule (using HB draws for first choice simulations) consistently beats the HBLC rule, and provides the lowest prediction errors (most accurate predictions) independent of which performance measure is considered. The RFC rule usually performs second-best (on par with HBLC for drill hammer and glass insurance), and noticeably better (even much better for summer tires and tablets) than the LC rule. This was unexpected in light of the findings of our Monte Carlo study. The FC rule performs either clearly the worst (drill hammer, lottery, tablet) or at least not (substantially) better than the LC rule (glass insurance, summer tires), which is in line with the findings of the Monte Carlo study.[22] While the HBFC rule only slightly outperforms the RFC rule in two data sets (lottery, tablets), it provides much more accurate shares of choice predictions in two other categories (glass insurance, summer tires). Further, the HBLC rule performs (clearly) worse than the RFC rule in three data sets (tablets, summer tires, lottery), and even somewhat worse than the LC rule for the lottery data.

Overall, using HB draws as inputs to first choice simulations (HBFC) performs extremely robustly in all five empirical data sets. It also seems to work better than the HBLC rule in empirical settings (as opposed to the results of the Monte Carlo study). Improvements from using the HBFC rule over the HBLC rule are especially noteworthy for summer tires and tablets. For summer tires, the HBFC rule provides an improvement in absolute errors (measured by RMSE and MAE) over the HBLC rule of nearly one percentage point.

Even an improvement in shares of choice predictions of less than one percentage point can have dramatic effects on revenues. Take for example the tablet market with its sales volume of 14 million units at the time the data were collected in September of 2010. The market share was 21.2% for one of the industry's leaders who offered its tablet for 600 euros. A prediction error of 0.5 share points corresponds to a revenue forecast that is off by 42 million euros in this case. Similarly, the summer tire market in Germany had a market volume of 17 million units in 2011. Considering one of the largest players in this market at that time with a market share of 30.5%, and a unit tire price of 66.59 euros, the same prediction error of 0.5 share points would generate a revenue forecast that is off by more than 5.5 million euros.[23] Our empirical findings regarding the comparative performance of the choice rules therefore clearly confirm the results of our Monte Carlo study, suggesting the use of HB draws and in particular the HBFC rule for preference simulations in practical settings.

As discussed in Section 2, Orme and Baker (2000) reported slightly more accurate shares of choice predictions based on the RFC rule as compared to the HBFC rule in their empirical CBC study. They argued with the *reverse number of levels (RNOL) effect* as a possible explanation for the slightly worse performance of the HBFC rule. According to the RNOL effect, within-respondent variances of HB draws for individual part-worths turn out higher for attributes with more levels when compared to attributes with fewer levels (which is reasonable from a statistical point of view, since levels of attributes with a larger number of levels occur less frequently in the choice task design). Since the variance of the part-worths is inversely related to their scaling, those scaling differences might have caused systematic biases in the choice shares predictions (Orme & Baker, 2000, p. 11). Based on our empirical findings, we don't think that the RNOL effect was the primary driver of their result, since in each of our five empirical data sets the number of levels across attributes varies to a greater or lesser extent (e.g. in the two categories of glass insurance and summer tires where the HBFC rule clearly outperformed the RFC rule, the number of attribute levels varies between two and six or three and ten across attributes). The slightly worse predictive performance of the HBFC rule in their study might instead have been caused by an insufficient number of burn-in iterations (35,000) and/or the specific characteristics of their empirical data set.[24] In addition, Orme and Baker (2000) only used 500 HB draws for shares of choice predictions for each respondent, while we used 1000 HB draws per respondent. Table 1 in the paper by Orme and Baker (2000, p. 8) reveals that the mean absolute error (MAE) between simulated and actual choice shares consistently decreased with an increasing number of HB draws per respondent, and it can be assumed that the prediction error in their study could have been further reduced if 1000 HB draws per respondent were used as well. As mentioned before in Section 2, Orme (2017) recently empirically compared shares of choice predictions based on the HBLC rule and the RFC rule in the course of Sawtooth Software's 2016 hosted modeling competition. Importantly, he reported slightly more accurate shares of choice predictions for the HBLC rule despite using only 200 draws per respondent, leaving much room for speculation what would happen for a much larger number of draws.

6. Summary, limitations, and outlook

The use of HB estimation techniques for CBC data offers the opportunity to directly use HB random draws for conjoint simulations. However, only few academic studies have been conducted that assess the predictive accuracy of preference simulations based on HB draws, and there is only one study that was dedicated to the comparison of choice share predictions based on HB draws with those based on the RFC rule (beyond the brief note by Orme (2017)). Furthermore, all previous studies were based on empirical data. The advantage of the artificial data we used in a first step is that experimental factors that may affect shares of choice predictions can be varied systematically, and undesirable confounding factors can be held constant. To the best of our knowledge, no study for CBC data exists that has systematically explored the shares of choice predictions based on HB draws in comparison to traditional choice rules like the first choice (FC), the logit choice (LC), or the randomized first choice (RFC) rules. The RFC rule was developed to correct for product similarity and consequently address the IIA property in preference simulations. Since the accuracy of choice share predictions can be different depending on the choice rule considered, this paper explored which choice rule should be used to predict preference shares as accurately as possible. From a theoretical perspective, if a Bayesian method is employed to estimate model parameters (like part-worth utilities), a fully Bayesian approach must be used to estimate quantities derived from this Bayesian model (like shares of choice predictions). The related objective of the current study therefore was to systematically explore whether the theoretically only correct way of using HB draws automatically coincides with more accurate shares of choice predictions, and the answer is a clear yes.

[22] At the individual holdout task level, the HBFC rule provides the best predictive performance in 78 out of 125 cases (5 data sets x 5 random holdouts x 5 performance measures = 125 performance evaluations), followed by the HBLC rule (17), the RFC rule (15), the LC rule (10), and the FC rule (5). The results for each single random holdout are available from the authors upon request.

[23] Processing times for model estimation ranged between seven minutes (430 seconds) for drill hammers and seven hours (25,670 seconds) for summer tires. Processing times for the choice rules were in all cases (i.e. for each data set and choice rule) lower than five seconds, and a maximum of 4.7 seconds for the tablet data with 1,000 respondents using the HBFC rule. Note that the number of respondents, and for summer tires in addition the number of parameters, were (much) larger than in the Monte Carlo study (compare Table 8).

[24] We used a minimum of 199,000 burn-in iterations throughout the paper (see Appendix A).

In particular, we investigated the capability of the different choice rules to handle nearly similar alternatives and assess how well the choice rules tolerate the IIA property. Not accounting for the IIA property in market simulations can lead to wrong managerial implications. Our Monte Carlo study was directed to analyze the conditions under which one of the choice rules recovers preference shares better than the other. We experimentally manipulated three factors (the number of choice tasks, the sample structure, as well as the number of parameters referred to as model complexity), and evaluated the performance of the choice rules under the different experimental conditions using several statistical criteria for predictive accuracy. Analyses of variance were conducted to assess the impact of the factors on the measures of predictive accuracy. To investigate how well the choice rules account for the IIA property, two different holdout choice scenarios containing three alternatives each were designed. Holdout tasks were designed to contain two extremely similar alternatives on the one hand, aiming to make predictions highly prone to the IIA property. Alternatively, holdout tasks were designed to contain alternatives that were different from each other in their profiles.

For both types of holdout choice scenarios, the results of the ANOVAs showed a statistically significant impact of both the number of choice tasks, sample structure, model complexity, and the type of choice rule on all measures of predictive accuracy. With regard to effect sizes, the type of choice rule in particular, the number of choice tasks, and the model complexity turned out to be primary drivers of shares of choice predictions. In general, the comparison between the two types of holdout choice scenarios revealed that prediction errors are higher when two very similar alternatives exist, because in these cases, predictions are more prone to the IIA property. As expected, a more complex model with many parameters to be estimated as well as lower numbers of choice tasks led to higher prediction errors. This is the case independent of the type of holdout choice scenario.

Concerning the choice rule, we found noticeable differences depending on the type of the holdout choice scenario. For holdouts containing two extremely similar alternatives, applying the FC rule led to the highest prediction errors. This is reasonable since the strictly deterministic FC rule tends to overestimate (underestimate) shares of choice for alternatives with marginal higher (lower) utilities, as can be expected for the two extremely similar alternatives. Accordingly, the results showed that the FC rule is more appropriate for predicting shares of choice when there is less influence of the IIA, i.e. when alternatives clearly differ from one other. In the latter case, much smaller differences between the FC, LC, and RFC rules were found.

With respect to the LC rule, the results clearly showed that accounting for heterogeneity by using individual-level utilities (estimated by HB) instead of aggregate-level utilities (estimated by the aggregate MNL) largely softened the IIA property, leading to much more accurate choice predictions. Although as expected, predictive validity measures were as a rule much worse for holdout choice scenarios with extremely similar alternatives when aggregate utilities were used, the performance measures were highly improved and nearly on par with both holdout choice scenarios after having accounted for heterogeneity through HB estimation.

The accuracy of choice share predictions based on the RFC rule was not superior to those based on the LC rule (applied to individual utilities) independent of the type of holdout tasks. This is particularly surprising for the holdout choice scenarios with two similar alternatives, because adding some random error to the point estimates of part-worths was expected to correct for product similarity, which causes similar alternatives to compete more closely with one another and to soften the IIA property as a result. Moreover, for holdout choice scenarios with different alternatives, the FC rule performed only slightly worse than RFC.

The most important finding of our Monte Carlo study was that using HB draws as inputs to either first choice simulations (HBFC) or logit choice simulations (HBLC) dramatically improved shares of choice predictions. While prediction errors from applying these two choice rules turned out much lower compared to all other choice rules independent of the type of holdout scenario, differences in the predictive performances between the two choice rules were only marginal and not significant for all measures of performance (with HBLC consistently performing slightly better). We did not expect such a clear superiority from using HB draws, in particular not when compared to the RFC rule. Contrary to the RFC rule, using draws directly from the HB estimation model preserves the individual respondent errors for shares of choice predictions. HB draws not only involve the Gumbel-distributed error of the MNL model, but additionally reflect a respondent's individual scaling depending on the quality of her/his answers along the choice tasks. This makes HB draws more powerful than the RFC rule, the latter adding a completely new model-independent attribute error to point estimates, and ignoring the individual scaling of respondents. Furthermore, compared to the LC rule applied to individual point estimates, it is clearly possible to balance out even more of the IIA bias via HB draws. Again, taking into account the uncertainty of the respondents along the choice tasks seems to make the difference here. So when HB is used for part-worth estimation, HB random draws should be saved and directly employed for preference simulations. Importantly, all five statistics we used for measuring predictive accuracy (MAE, MSE, RMSE, MAPE, and RAE) indicated the clear superiority of using HB random draws for preference simulations. Altogether, the current paper presents clear evidence that using HB draws not only provides the most accurate shares of choice predictions but, more importantly, is theoretically correct.

As a refinement, we further controlled for potentially confounding parameter recovery effects on shares of choice predictions. Although a bad parameter recovery as suspected increased prediction errors, the accuracy of shares of choice predictions remained most strongly influenced by the type of choice rule used. Finally, computation times for shares of choice predictions were within one second for all choice rules, including HBFC and HBLC, meaning that the processing times of the HBFC and HBLC rules are economically comparable to the other choice rules, with much more accurate shares of the choice predictions they achieve.

To check the validity and generalizability of our results obtained from artificial data for real data settings, we compared the performance of the choice rules with regard to shares of choice predictions in a second step based on five empirical data sets. The results here confirmed most of the findings of the Monte Carlo study, specifically regarding how the use of HB draws was clearly preferable for choice share predictions. However, the HBFC rule (i.e. combining HB draws with first choice simulations) outperformed the HBLC rule, providing the lowest prediction errors in these empirical settings. Further, the RFC rule usually performed second best (sometimes on par with HBLC) and therefore better than in the Monte Carlo study. The finding that the HBFC rule provided the lowest prediction errors across all five empirical data sets is an important preliminary result for practitioners who are concerned with choice-based conjoint analysis for preference simulations. Processing times for the HBFC rule did not exceed five seconds despite a (much) higher number of respondents in all five empirical data sets (up to 1000), as well as a much higher model complexity (94 parameters per respondent) for the summer tire data compared to the simulated data of the Monte Carlo study.

Our results further indicated that the composition of alternatives in holdout tasks, i.e. whether holdouts contain highly similar alternatives or not, has an influence on prediction accuracy. This is an important issue because the internal validity in em-

pirical CBC studies is often assessed in terms of fixed holdout tasks. However, the accuracy of choice share predictions can be quite different depending on the chosen constellation of alternatives in the fixed holdout tasks. Hence it seems more reasonable to use random choice tasks as holdouts as we did in our empirical study.

The limitations of our study should be addressed in future research. First, it is unclear why the RFC rule did not outperform the LC rule in our Monte Carlo study, while it consistently did so in our empirical study. One idea could be to incorporate simplification strategies of respondents as additional experimental factor into Monte Carlo studies (e.g. in terms of certain percentages of respondents who make their choices based only on key attributes like brand or price) to capture behavioral effects. Additionally, other behavioral patterns could be considered, including boredom, fatigue, or even some kind of cheating behavior by human respondents that lead to simplification patterns as well and reduce the data quality. As the RFC rule captures attribute variability and hence uncertainty about respondents' preferences, this might provide an explanation for the better performance of the RFC rule in empirical studies. Second, more research is needed to precisely explore what really drives the much better predictive performance of the HBFC rule compared to the RFC rule, as observed in both our Monte Carlo and empirical studies. Third, more research is also needed to explore why the HBLC rule seems to perform less robustly in empirical settings (it performed worse than the RFC rule in three data sets and even worse than the LC rule in one data set), while the HBFC rule performed extremely well under both simulated and empirical conditions. One explanation for this may be that the HBLC rule further distributes a respondent's preference uncertainty contained in her/his draws among the alternatives according to a probability distribution, which might be harmful to adequately capturing minority preferences in data. With this in mind, consider the following example: It is well-known that the vast majority of car drivers prefer front-wheel drive to rear-wheel drive. Looking only at point estimates from the HB model for a rare respondent who actually prefers rear-wheel drive would in many situations most likely obscure her/his true preference due to the shrinkage effect of the HB model to the population mean (and hence to front-wheel drive). However, the respondent's preferences for rear-wheel drive would become visible within a part of the individual HB draws, and simulating first choices based on these HB draws would result in a certain proportion of choosing rear-wheel drive. Applying HBLC instead of HBFC could here dilute this proportion again to some extent because the first choices in favor of rear-wheel drive would be further distributed among alternatives according to the draw-specific logit choice probabilities. If this were true, a larger number of choice tasks should improve shares of choice predictions based on the HBLC rule; in this case, more individual respondent information would be available decreasing the effect of pooling and eventually allowing a larger deviation of an individual respondent's part-worths from the population means. Since this effect is not clear from Table 9, it should be analyzed in future research via a simulation exercise with more and less pooling in the Wishart distribution as experimental factor. Regardless of this possible explanation, the conditions that could drive a better performance of the HBFC rule compared to the HBLC rule in empirical applications (as opposed to the results of our Monte Carlo study) should be systematically explored in the future. Fourth, since the HBLC rule performed slightly better than the HBFC rule in our Monte Carlo study, future studies could analyze the conditions under which the HBFC rule approximates HBLC probabilities well versus less well. One hypothesis is that the HBFC rule may have greater difficulties in the approximation of the HBLC probabilities when there are larger differences in utilities between alternatives. And finally, it would be interesting to assess whether the superiority of using HB draws for share predictions is preserved with respect to the external validity of CBC studies. Because this requires the availability of/access to real market shares, we leave this issue to future research.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2021.05.056.

Appendix A. The HB mixed logit model for choice-based conjoint analysis

The HB mixed logit model is a random effects model accommodating heterogeneity of respondents by assuming a specific probability distribution for individual-level parameters. In contrast to classical random effects models where only the parameters of the underlying probability distribution can be estimated, the hierarchical structure of HB random effects models allows inferences to be drawn regarding individual-level parameters by combining the information from the assumed probability distribution of all individuals, and each individual's choice data (e.g., Allenby & Ginter, 1995; Rossi & Allenby, 1993; Rossi, Allenby & McCulloch, 2005).

Individuals' choices are captured by an MNL model on the lower level of the HB mixed logit. Accordingly, the choice probability $P_n(j)$ of individual n to choose alternative j from a certain choice task is:

$$P_n(j) = \frac{\exp\left(\beta_n x_j\right)}{\sum_{m=1}^{J} \beta_n x_m},$$

where β_n denotes the vector of part-worths of individual n, and x_j is a binary vector coding the presence (=1) or absence (=0) of attribute levels for alternative j. On the upper level of the HB mixed logit, we use the multivariate normal distribution as population distribution (first-stage prior distribution):

$$\beta_n \sim \mathrm{N}\left(\overline{\beta}, V_\beta\right),$$

where the covariance matrix V_β captures the degree of heterogeneity across individuals and the covariances between the respondents' part-worths. $\overline{\beta}$ is the expectation of the normal distribution and here represents the vector of means of the distribution of individuals' part-worths.

For full Bayesian inference, we assign (as usual) a normal distribution for the vector of means $\overline{\beta}$ and an inverse Wishart distribution for the covariance matrix V_β as hyperprior distributions to estimate the parameters of the prior distribution (cf. Rossi et al., 2005). Random draws from the posterior distribution which is not analytically tractable (e.g., Allenby et al., 1995) are generated by applying Markov chain Monte Carlo (MCMC) techniques. After convergence of the Markov chain (i.e., after the burn-in phase), the generated draws are typically averaged to obtain point estimates of the parameters. These are used as inputs for preference simulations based on the FC and LC rules, and also build the starting point for the RFC rule (where new draws are generated around these point estimates). Alternatively, the individual HB draws can be directly used for preference simulations (referred to as HBFC and HBLC rules). For our Monte Carlo study, we saved 1000 draws per respondent, resulting for 200 respondents in a total of 200,000 sets of twelve part-worths for simple models (six attributes, three levels), and 48 part-worths for complex models (twelve attributes, five levels) for each treatment and replication.[25] As an example, for our 1000 respondents in our empirical data set for tablets,

[25] Note that for H attributes with I levels each, H times (I-1) part-worths need to be estimated independent of whether a dummy- or effects-coding is used.

1,000,000 sets of 43 part-worths resulted. We saved only every fifth draw after convergence to reduce the amount of correlation across the draws (i.e. we always ran 5000 iterations after convergence). Our default setting for the burn-in phase was 199,000 iterations, but in many cases the Markov chains needed more iterations to converge based on the PSRF measure used for the convergence diagnostic (see Supplementary Materials).[26]

Appendix B. Distributions of individual-level part-worths

Recall that, following the data generation process by Wirth (2010), for 80% of the attribute levels the corresponding population mean betas (represented by the black circles on the bars in Fig. B.1) were randomly drawn from the range between −2 and 2 (displayed by the dashed lines), and for each 10% of the attribute levels the corresponding population mean betas were drawn from the intervals [−5, −2] and [2, 5]. Normally distributed preferences around each mean beta were generated subsequently, as exemplarily illustrated by the normal density plot for one of the attribute levels in the figure for a less heterogeneous sample. Note that the variances of the normal distributions of the attribute level part-worths were allowed to vary, as represented by the different lengths of the bars. As a result, a unique multivariate normal distribution of preference heterogeneity was generated for each run in the Monte Carlo study (compare Section 4.2).

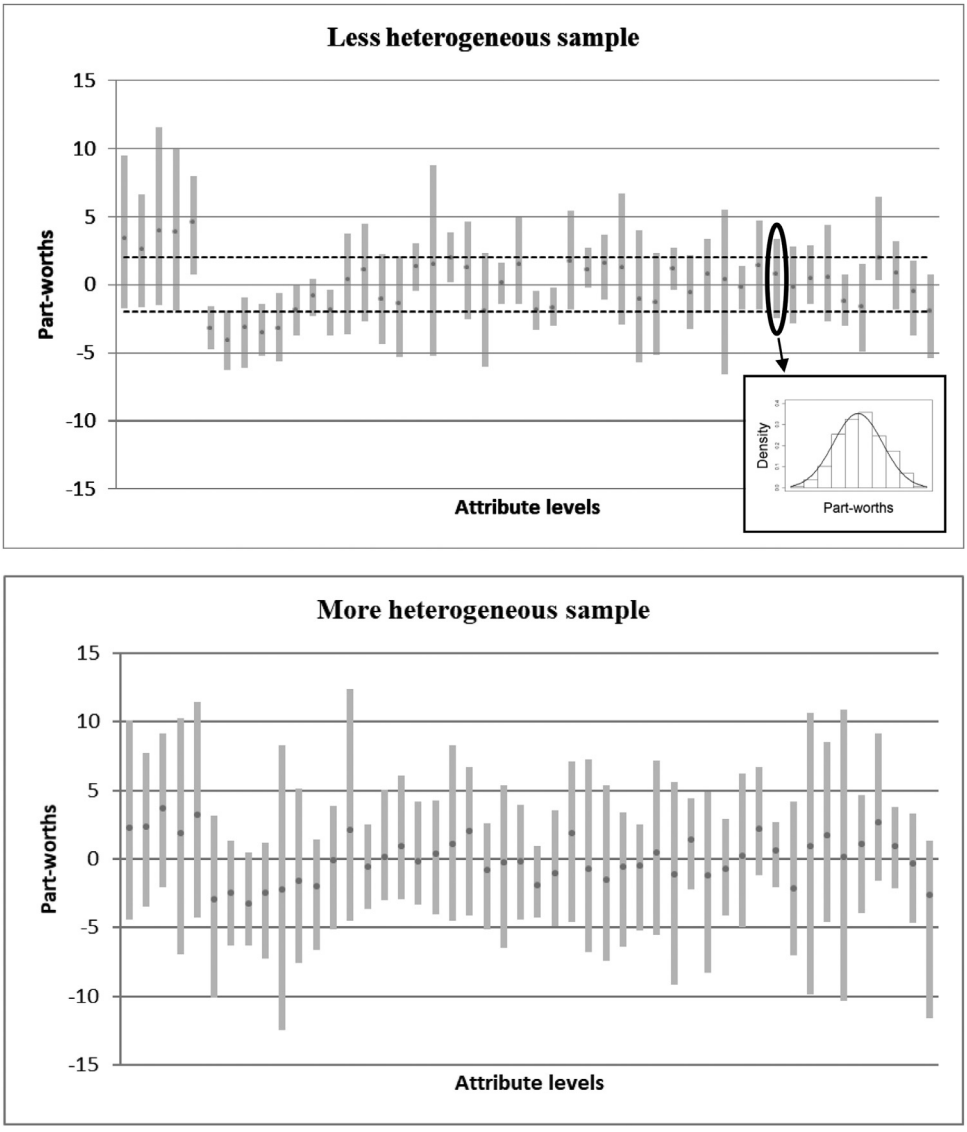

Fig. B.1. Distributions of individual-level part-worths for a less heterogeneous sample (upper figure) and for a more heterogeneous sample (lower figure), exemplarily for a treatment with twelve attributes and five attribute levels. The length of a bar corresponds to the range of the generated true part-worths for a respective attribute level and hence reflects the amount of heterogeneity across respondents.

[26] 199,000 iterations (11%), 299,000 iterations (57%), 399,000 iterations (29%), 499,000 iterations (3%).

Appendix C. Means of the five measures of predictive accuracy (including predictions from the aggregate MNL model)

This appendix contains the means of performance measures by experimental condition for each factor level including the shares of choice predictions from the aggregate MNL model. See Table C.1 for holdout choice scenarios where two out of three alternatives are extremely SIMILAR, and Table C.2 for holdout choice scenarios with DISSIMILAR alternatives.

Table C1
Means of performance measures by experimental condition (for each factor level)[a] for holdout choice scenarios where two out of three alternatives are extremely SIMILAR (including the aggregate MNL model).

	Predictive Accuracy				
Factor	MAE	MSE	RMSE	MAPE	RAE
Choice rule					
(1) FC	$11.134^{4,5,6**}$	$190.009^{4,5*,6**}$	$12.274^{4,5,6**}$	$36.048^{4,5,6**}$	$.337^{4,5,6**}$
(2) LC	$9.129^{4,5,6**}$	131.000^{6**}	$10.187^{4,5,6**}$	$30.414^{4,5,6**}$	$.277^{4,5,6**}$
(3) RFC	$9.759^{4,5,6**}$	150.631^{6**}	$10.802^{4,5,6**}$	$31.829^{4,5,6**}$	$.296^{4,5,6**}$
(4) HBFC	$5.212^{1,2,3,6**}$	$49.447^{1*,6**}$	$5.860^{1,2,3,6**}$	$16.395^{1,2,3,6**}$	$.158^{1,2,3,6**}$
(5) HBLC	$4.825^{1,2,3,6**}$	$43.487^{1*,6**}$	$5.425^{1,2,3,6**}$	$15.304^{1,2,3,6**}$	$.147^{1,2,3,6**}$
(6) Aggregate MNL	**$23.258^{1,2,3,4,5**}$**	**$727.534^{1,2,3,4,5**}$**	**$25.438^{1,2,3,4,5**}$**	**$71.310^{1,2,3,4,5**}$**	**$.705^{1,2,3,4,5**}$**
Number of choice tasks					
7	12.250^{15**}	272.483	13.542^{15**}	39.265^{15**}	$.371^{15**}$
11	10.665	197.429	11.836	33.998	.323
15	8.743^{7**}	176.142	9.615^{7**}	27.388^{7**}	$.265^{7**}$
Sample structure					
Homogeneous	11.292	238.027	12.476	36.104^{*}	.342
Heterogeneous	9.813	192.676	10.853	30.996^{*}	.297
Model complexity					
Simple	9.089^{**}	185.379	10.052^{**}	29.215^{**}	$.276^{**}$
Complex	12.016^{**}	245.324	13.276^{**}	37.886^{**}	$.364^{**}$

*Indicates that the difference between two means is significant at the .05 level (as indicated by the Bonferroni correction method).
**Indicates that the difference between two means is significant at the .01 level (as indicated by the Bonferroni correction method).
[a]Superscripts on means refer to the factor levels and indicate the factor level that leads to a significant difference between the means of the measure of performance.

Table C2
Means of performance measures by experimental condition (i.e. for each factor level)[a] for holdout choice scenarios with DISSIMILAR alternatives (including the aggregate MNL model).

	Predictive Accuracy				
Factor	MAE	MSE	RMSE	MAPE	RAE
Choice rule					
(1) FC	$9.986^{4,5,6**}$	154.685^{6**}	$11.096^{4,5,6**}$	46.370^{6**}	$.303^{4,5,6**}$
(2) LC	$9.067^{4,5,6**}$	132.377^{6**}	$10.077^{4,5,6**}$	42.889^{6**}	$.275^{4,5,6**}$
(3) RFC	$9.253^{4,5,6**}$	137.475^{6**}	$10.290^{4,5,6**}$	43.476^{6**}	$.280^{4,5,6**}$
(4) HBFC	$4.386^{1,2,3,6*}$	37.117^{6**}	$4.874^{1,2,3,6**}$	22.472^{6**}	$.133^{1,2,3,6**}$
(5) HBLC	$4.136^{1,2,3,6*}$	33.698^{6**}	$4.605^{1,2,3,6**}$	21.688^{6**}	$.126^{1,2,3,6**}$
(6) Aggregate MNL	**$18.315^{1,2,3,4,5**}$**	**$575.989^{1,2,3,4,5**}$**	**$20.583^{1,2,3,4,5**}$**	**$124.512^{1,2,3,4,5**}$**	**$.555^{1,2,3,4,5**}$**
Number of choice tasks					
7	10.663^{15*}	218.884	11.874^{15*}	64.355	$.323^{15*}$
11	8.739	143.023	9.775	45.163	.265
15	8.169^{7*}	173.764	9.113^{7*}	41.185	$.248^{7*}$
Sample structure					
Homogeneous	9.443	176.083	10.519	48.306	.286
Heterogeneous	8.937	181.030	9.989	52.162	.271
Model complexity					
Simple	8.115^{**}	163.896	9.042^{**}	51.622	$.246^{**}$
Complex	10.266^{**}	193.218	11.466^{**}	48.847	$.311^{**}$

*Indicates that the difference between two means is significant at the .05 level (as indicated by the Bonferroni correction method).
**Indicates that the difference between two means is significant at the .01 level (as indicated by the Bonferroni correction method).
[a]Superscripts on means refer to the factor levels and indicate the factor level that leads to a significant difference between the means of the measure of performance.

Appendix D. Parameter recovery and shares of choice predictions

Section 4.4 showed that the performance of shares of choice predictions in particular depends on the type of choice rule used (large effect sizes), as well as number of choice tasks and model complexity (medium effect sizes). On the other hand, we did not find significant (27 out of 30) or substantial (3 out of 30) interaction effects between the type of choice rule and the manipulated conditions, implying that the differences in the performances of the choice rules are hardly or not at all influenced by the experimental conditions. However, the shares of choice predictions might generally have been affected by the goodness of parameter recovery, i.e. how well the generated "true" part-worth utilities that served as input to the choice rules could be recovered in the parameter estimation stage under the different experimental conditions.

Table D.1
Means of RMSE (β) by experimental condition (for each factor level)[a].

	Parameter Recovery
Experimental Condition	**RMSE (β)**
Number of choice tasks	
7	$4.190^{11,15**}$
11	3.480^{7**}
15	3.019^{7**}
Sample structure	
Less heterogeneous	4.218**
More heterogeneous	2.908**
Model complexity	
Simple	2.115**
Complex	5.011**

*Indicates that the difference between two means is significant at the 0.05 level (as indicated by the Bonferroni correction method).
**Indicates that the difference between two means is significant at the 0.01 level (as indicated by the Bonferroni correction method).
[a]Superscripts on means refer to the factor levels and indicate the factor level that leads to a significant difference between the means of the measure of performance.

Parameter recovery is usually measured by the root mean square error (RMSE) between the generated "true" part-worths and the re-estimated part-worths (see for example the aforementioned Monte Carlo studies of Andrews, Ainslie & Currim, 2002, Andrews, Ansari, & Currim, 2002; Vriens et al., 1996):

$$\text{RMSE}(\beta) = \sqrt{\frac{\sum_n \sum_h \sum_i (\hat{\beta}_{nhi} - \beta_{nhi})^2}{NHI}}$$

where N, H, and I respectively denote the numbers of respondents, attributes and attribute levels. Table D.1 summarizes the means of the RMSE (β) for parameter recovery depending on the experimental condition. Note that parameter recovery is independent of the type of holdout choice scenario (SIMILAR, DISSIMILAR), since the construction of holdouts does not affect the preference estimation stage.

The results in Table D.1 reveal a significant impact of all factors on the parameter recovery (RMSE (β)), which becomes considerably worse when the CBC model is complex (twelve attributes, five attribute levels) rather than simple (six attributes, three attribute levels). Analogous to our findings on prediction errors, parameter recovery significantly benefits from a higher amount of respondent heterogeneity. Further, parameter recovery turns out significantly worse when the number of choice tasks is low (seven).

We ran five regressions to assess the impact of the parameter recovery on shares of choice predictions, each time using one of the five measures of predictive accuracy (RMSE, MSE, MAE, MAPE, RAE) as the dependent variable, and the RMSE (β) for parameter recovery, the different choice rules, and the type of holdout choice scenario as independent variables. Each regression was based on 720 observations (twelve treatments, six runs, two holdout choice scenarios, five choice rules) with 713 degrees of freedom for error. Table D.2 displays the results for the regression models (compare Section 4.3 for the measures of predictive accuracy).

Note that the holdout choice scenario with dissimilar alternatives and the HBLC rule were coded as reference levels. As can be seen from Table D.2, the prediction error significantly depends on the parameter recovery ($p<.0001$) for all five measures of predictive accuracy. The following discussion refers to the prediction RMSE and is representative in all aspects for the other four measures of predictive accuracy.

Table D2
Regression coefficients and model fit.

Dependent variable	**RMSE**		**MSE**		**MAE**		**MAPE**		**RAE**	
	Coef.[a]	t value[b]	Coef.[a]	t value[b]	Coef.[a]	t value[b]	Coef.[a]	t value[b]	Coef.[a]	t value[b]
Intercept[c]	.490	.897	-61.647	-4.560	.423	.847	9.003	5.467	.013	.879
	(.547)	(.370)	(13.519)	(<.0001)	(.499)	(.397)	(1.647)	(<.0001)	(.015)	(.380)
RMSE (β)[d]	1.169	11.855	26.191	10.748	1.048	11.650	3.981	13.407	.032	11.643
	(.099)	(<.0001)	(2.437)	(<.0001)	(.090)	(<.0001)	(.297)	(<.0001)	(.003)	(<.0001)
FC rule	6.670	12.327	133.755	9.999	6.080	12.311	22.714	13.937	.184	12.272
	(.541)	(<.0001)	(13.377)	(<.0001)	(.494)	(<.0001)	(1.630)	(<.0001)	(.015)	(<.0001)
RFC rule	5.531	10.221	105.460	7.884	5.026	10.177	19.157	11.754	.152	10.140
	(.541)	(<.0001)	(13.377)	(<.0001)	(.494)	(<.0001)	(1.630)	(<.0001)	(.015)	(<.0001)
LC rule	5.117	9.456	93.096	6.960	4.617	9.350	18.156	11.140	.139	9.316
	(.541)	(<.0001)	(13.377)	(<.0001)	(.494)	(<.0001)	(1.630)	(<.0001)	(.015)	(<.0001)
HBFC rule	.352	.651	4.689	.351	.318	.644	.938	.575	.009	.611
	(.541)	(.515)	(13.377)	(.726)	(.494)	(.520)	(1.630)	(.565)	(.015)	(.542)
Type SIMILAR	.721	2.107	13.845	1.636	.646	2.070	-9.381	-9.101	.020	2.075
	(.342)	(.035)	(8.460)	(.102)	(.312)	(.039)	(1.031)	(<.0001)	(.009)	(.038)
R^2 (Corrected R^2)	.366 (.360)		.286 (.280)		.362 (.356)		.464 (.459)		.361 (.356)	

[a]Std. error in parentheses.
[b]p value in parentheses.
[c]The intercept refers to the holdout choice scenario of type DISSIMILAR and the HBLC rule as a reference category.
[d]Parameter recovery.

On average, if the parameter RMSE increases (worsens) by one unit, the prediction RMSE also increases (worsens) by about one unit ($\gamma_{\text{RMSE}(\beta)} = 1.169$). However, note that the correlation coefficient between parameter RMSE and prediction RMSE is only 0.184. Dropping the choice rules as predictors from the regression model would dramatically decrease the R square from 0.366 to 0.129, which underlines the importance of using the right choice rule for shares of choice predictions. Note also that the parameter estimates and significance levels for the different choice rules resemble our ANOVA findings with regard to the effects of the choice rules on prediction errors measured in terms of RMSE (compare Tables 6 and 7). In particular, the HBFC rule does not perform significantly worse compared to the HBLC rule (in contrast to the other choice rules). On the other hand, although considering the experimental conditions (number of choice tasks, sample structure, model complexity) as additional predictors in the regression model would increase the R square by 0.121 (from 0.366 to 0.487), it would cause serious multicollinearity problems, as reflected by a wrong sign for the coefficient of the parameter RMSE ($\gamma_{\text{RMSE}(\beta)} = -1.873$, $p<.0001$). In fact, the impact of differences in the parameter recovery by experimental condition on shares of choice predictions is captured to a greater extent by including the parameter recovery RMSE as a predictor (also compare Table D.1 combined with Tables 6 and 7).[27] Overall, parameter recovery has an influence on prediction errors (as expected), but the accuracy of shares of choice predictions is nevertheless most influenced by the type of choice rule used. Finally, note that parameter recovery can only be evaluated in Monte Carlo studies when "true" parameters are known, but not in empirical studies.

Appendix E. Computing times

Parameter estimation and shares of choice predictions were performed on a Windows 7 computer with a 3.5 GHz 12 core i7 processor. The platform for estimating the HB mixed logit model was the CBC/HB module of Sawtooth Software (version 5.5.3). We used the R package (version 3.5.1) as the platform for data generation, implementation of the choice rules (except for the RFC rule where we used an Excel plug-in made available by Sawtooth), and calculation of the performance measures. Note that all choice rules (FC, LC, RFC, HBFC, HBLC) shared the same individual-level part-worth estimates depending on the experimental condition and individual run (replication) per treatment. Table E.1 displays the mean run times and the span of individual runtimes for 1000 iterations of the MCMC sampler by factor level (parameter estimation stage) as well as the corresponding mean run times and the span of individual runtimes for the different choice rules by experimental condition (preference simulation stage). Note that shares of choice predictions using the FC rule and the LC rule were based on point estimates for each respondent, while RFC, HBFC, and HBLC simulations were based on 1000 (HB) draws per respondent. We report the runtimes per 1000 iterations for the model estimation stage because only a part of the HB models converged after 199,000 burn-in iterations (our default setting), compare Appendix A. In addition, we also provide the mean total runtimes and the span of individual total runtimes by factor level for estimating the HB Mixed Logit.

Computing times for model estimation are mostly affected by the model complexity, and also somewhat increase with the number of choice tasks. As expected, the amount of heterogeneity has virtually no impact on runtimes. The longest estimation time of 16.21 s per 1000 MCMC iterations was observed for a treatment with 15 choice tasks, a more heterogeneous sample structure, and high model complexity (twelve attributes, five levels), for which our default setting of 199,000 burn-in iterations to reach convergence of the Markov chain was sufficient. The longest total runtime of 6503 s or 108 min for model estimation was observed for a different treatment with 15 choice tasks, a less heterogeneous sample structure, and high model complexity (twelve attributes, five levels), where 399,000 burn-in iterations were needed for the HB model to converge.[28]

With regard to shares of choice predictions (preference simulation stage), we observe mean runtimes below one second for each of the five choice rules independent of the experimental condition. Given a minimum total runtime of 445 s for model estimation (see Table E.1, parameter estimation stage), the runtimes for

Table E.1
Computing times [*sec*] for (1) estimating the HB mixed logit model by factor level (parameter estimation stage), and for (2) shares of choice predictions by choice rule (preference simulation stage). The span of individual run times is in parentheses.

Factor	Parameter estimation stage		Preference simulation stage				
	Time [*sec*] per 1000 iterations	Total run times [*sec*]	Time [*sec*]				
			FC	LC	RFC	HBFC	HBLC
Number of choice tasks							
(1) 7	6.80	2289.00	.29	.12	.92	.90	.28
	[1.46,15.38]	[445,5992]	[.25,.33]	[.09,.14]	[.78,1.00]	[.86,.94]	[.23,.31]
(2) 11	7.87	2337.79	.31	.12	.92	.90	.28
	[1.83,15.02]	[557,5496]	[.25,.34]	[.09,.14]	[.78,1.00]	[.84,.98]	[.23,.31]
(3) 15	8.05	2664.29	.32	.13	.89	.93	.30
	[2.24,16.21]	[680,6503]	[.23,.39]	[.09,.19]	[.78,1.00]	[.84,1.00]	[.23,.36]
Sample structure							
(1) Less heterogeneous	7.62	2735.39	.30	.12	.93	.91	.28
	[1.48,16.10]	[449,6503]	[.23,.34]	[.09,.14]	[.79,1.00]	[.84,.98]	[.23,.31]
(2) More heterogeneous	7.52	2125.33	.31	.13	.89	.91	.29
	[1.46,16.21]	[445,4844]	[.25,.39]	[.09,.19]	[.78,1.00]	[.84,1.00]	[.23,.36]
Model complexity							
(1) Simple	1.92	630.25	.31	.12	.83	.92	.30
	[1.46,2.43]	[445,958]	[.30,.33]	[.11,.14]	[.78,1.00]	[.89,.98]	[.24,.31]
(2) Complex	13.22	4230.47	.30	.12	.99	.90	.27
	[10.58,16.21]	[2159,6503]	[.23,.39]	[.09,.19]	[.99,1.00]	[.84,1.00]	[.23.36]

[27] For example, the pairwise correlation between the parameter RMSE (RMSE (β)) and the factor model complexity is already 0.85. Detailed results for the regression models excluding the choice rules, and the regression models including the dummies for the experimental conditions as additional predictors, are available from the authors upon request.

[28] Note that the lower and upper bounds of the ranges of the computing times for runtimes per 1,000 iterations (second column of Table E.1) and total runtimes (third column of Table E.1) may not coincide with the same treatment, since the number of burn-in iterations to achieve convergence can differ between treatments.

all five choice rules are marginal, making the differences in runtimes between the choice rules seem negligible. Note that logit choice simulations are in general somewhat faster than first choice simulations (LC vs. FC, or HBLC vs. HBFC/RFC), and runtimes of the HBFC rule and the RFC rule are nearly identical. Overall, the processing times for both the HBFC rule (approximately 0.9 s) and the HBLC rule (approximately 0.3 s) are highly acceptable economically, given their much more accurate shares of choice predictions they achieved.

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*(422), 669–679.
Allenby, G. M., Arora, N., & Ginter, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research, 32*(2), 152–162.
Allenby, G. M., & Ginter, J. L. (1995). Using extremes to design products and segment markets. *Journal of Marketing Research, 32*(4), 392–403.
Alvarez, R. M., & Nagler, J. (1998). When politics and models collide: Estimating models of multiparty elections. *American Journal of Political Science, 42*(1), 55–96.
Andrews, R. L., Ainslie, A., & Currim, I. S. (2002). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *Journal of Marketing Research, 39*(4), 479–487.
Andrews, R. L., Ansari, A., & Currim, I. S. (2002). Hierarchical Bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. *Journal of Marketing Research, 39*(1), 87–98.
Arenoe, B. (2003). Determinants of external validity in CBC. In *Proceedings of the 10th Sawtooth software conference* (pp. 217–232).
Aribarg, A., Arora, N., & Kang, M. Y. (2010). Predicting joint choice using individual data. *Marketing Science, 29*(1), 139–157.
Baier, D., & Polasek, W. (2003). Market simulation using Bayesian procedures in conjoint analysis. In M. Schwaiger, & O. Opitz (Eds.), *Exploratory data analysis in empirical research* (pp. 413–421). Berlin: Springer.
Belloni, A., Freund, R., Selove, M., & Simester, D. (2008). Optimizing product line designs: Efficient methods and comparisons. *Management Science, 54*(9), 1544–1552.
Ben-Akiva, M. E., & Lerman, S. R. (1985). *Discrete choice analysis. Theory and application to travel demand*. Cambridge MA: MIT Press (MIT Press series in transportation studies, 9).
Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25*(2), 290–302.
Braun, A., Schmeiser, H., & Schreiber, F. (2016). On consumer preferences and the willingness to pay for term life insurance. *European Journal of Operational Research, 253*(3), 761–776.
Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455.
Camm, J. D., Cochran, J. J., Curry, D. J., & Kannan, S. (2006). Conjoint optimization: An exact branch-and-bound algorithm for the share-of-choice problem. *Management Science, 52*(3), 435–447.
Chakraborty, G., Ball, D., Gaeth, G. J., & Jun, S. (2002). The ability of ratings and choice conjoint to predict market shares. A Monte Carlo simulation. *Journal of Business Research, 55*(3), 237–249.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Debreu, G. (1960). Review of R. D. Luce, individual choice behavior: A theoretical analysis. *American Economic Review, 50*, 186–188.
Dotson, J. P., Howell, J. R., Brazell, J. D., Otter, T., Lenk, P. J., MacEachern, S., et al. (2018). A Probit model with structured covariance for similarity effects and source of volume calculations. *Journal of Marketing Research, 55*(1), 35–47.
Dotson, M. R., Büschken, J., & Allenby, G. M. (2020). Explaining preference heterogeneity with mixed membership modeling. *Marketing Science, 39*(2), 407–426.
Elrod, T., & Kumar, S. K. (1989). Bias in the first choice rule for predicting share. In *Proceedings of the 3rd Sawtooth software conference* (pp. 259–271).
Fahrmeir, L., & Kneib, T. (2011). *Bayesian smoothing and regression for longitudinal, spatial and event history data*. Oxford: Oxford University Press.
Feit, E. M., Beltramo, M. A., & Feinberg, F. M. (2010). Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Science, 56*(5), 785–800.
Finkbeiner, C. T. (1988). Comparison of conjoint choice simulators. In *Proceedings of the 2nd Sawtooth software conference* (pp. 75–103).
Foster, G., Turner, C., Ferguson, S., & Donndelinger, J. (2014). Creating targeted initial populations for genetic product searches in heterogeneous markets. *Engineering Optimization, 46*(12), 1729–1747.
Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472.
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. *European Journal of Operational Research, 219*(2), 368–378.
Gilbride, T. J., Lenk, P. J., & Brazell, J. D. (2008). Market share constraints and the loss function in choice-based conjoint analysis. *Marketing Science, 27*(6), 995–1011.
Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237–288.
Green, P. E., & Krieger, A. M. (1988). Choice rules and sensitivity analysis in choice simulators. *Journal of the Academy of Marketing Science, 16*(1), 114–127.
Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research, 8*(3), 355–363.
Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research, 5*(2), 103–123.
Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing, 54*(4), 3–19.
Halme, M., & Kallio, M. (2011). Estimation methods for choice-based conjoint analysis of consumer preferences. *European Journal of Operational Research, 214*(1), 160–167.
Halme, M., & Kallio, M. (2014). Likelihood estimation of consumer preferences in choice-based conjoint analysis. *European Journal of Operational Research, 239*(2), 556–564.
Hoogerbrugge, M., & van der Wagt, K. (2006). How many choice tasks should we ask? In *Proceedings of the 12th Sawtooth software conference* (pp. 97–110).
Huber, J., Orme, B., & Miller, R. (1999). Dealing with product similarity in conjoint simulations. *Sawtooth software research paper series*. Sequim, WA: Sawtooth Software.
Huber, J., Orme, B., & Miller, R. (2007). Dealing with product similarity in conjoint simulations. In A. Gustafsson, & F. Huber (Eds.), *Conjoint measurement: Methods and applications* (pp. 347–362). Berlin: Springer.
Jain, D. C., & Bass, F. M. (1989). Effect of choice set size on choice probabilities: An extended logit model. *International Journal of Research in Marketing, 6*(1), 1–11.
Karniouchina, E. V., Moore, W. L., van der Rhee, B., & Verma, R. (2009). Issues in the use of ratings-based versus choice-based conjoint analysis in operations management research. *European Journal of Operational Research, 197*(1), 340–348.
Kurz, P., & Binner, S. (2012). The individual choice task threshold: Need for variable number of choice tasks. In *Proceedings of the 16th Sawtooth software conference* (pp. 111–127).
Leeflang, P. S. H., Wittink, D. R., Wedel, M., & Naert, P. A. (2000). *Building models for marketing decisions*. Boston: Kluwer Academic Publishers.
Lenk, P. J., Desarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science, 15*(2), 173–191.
Louviere, J. J., Hensher, D. A., Swait, J. D., & Adamowicz, W. (2010). *Stated choice methods. Analysis and applications*. Cambridge: Cambridge University Press.
Louviere, J. J., & Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research, 20*(4), 350–367.
Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.
Maldonado, S., Montoya, R., & Weber, R. (2015). Advanced conjoint analysis using feature selection via support vector machines. *European Journal of Operational Research, 241*(2), 564–574.
McCullough, D. (2002). A user's guide to conjoint analysis. *Marketing Research, 14*(2), 19–23.
McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). New York: Academic Press.
McFadden, D., Train, K., & Tye, W. B. (1977). An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model. *Transportation Research Record, 637*, 39–46.
Meeran, S., Jahanbin, S., Goodwin, P., & Neto, J. Q. F. (2016). When do changes in consumer preferences make forecasts from choice-based conjoint models unreliable? *European Journal of Operational Research, 258*(2), 512–524.
Moore, W. L. (2004). A cross-validity comparison of rating-based and choice-based conjoint analysis models. *International Journal of Research in Marketing, 21*(3), 299–312.
Moore, W. L., Gray-Lee, J., & Louviere, J. J. (1998). A cross-validity comparison of conjoint analysis and choice models at different levels of aggregation. *Marketing Letters, 9*(2), 195–207.
Natter, M., & Feurstein, M. (2002). Real world performance of choice-based conjoint models. *European Journal of Operational Research, 137*(2), 448–458.
Orme, B. (1998). The benefits of accounting for respondent heterogeneity in choice modeling. *Sawtooth software research paper series*. Sequim, WA: Sawtooth Software.
Orme, B. (2017). Findings of the 2016 Sawtooth Software CBC modeling prize competition. In *Proceedings of the 19th Sawtooth software conference* (pp. 37–52).
Orme, B., & Baker, G. (2000). Comparing hierarchical Bayes draws and randomized first choice for conjoint simulations. In *Proceedings of the 8th Sawtooth software conference* (pp. 239–254).
Orme, B., & Huber, J. (2000). Improving the value of conjoint simulations. *Marketing Research, 12*(4), 12–20.
Orme, B., & Johnson, R. (2006). External effect adjustments in conjoint analysis. *Sawtooth software research paper series*. Sequim, WA: Sawtooth Software.
R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rao, V. R. (2014). *Applied conjoint analysis*. New York NY: Springer.
Ray, P. (1973). Independence of irrelevant alternatives. *Econometrica, 41*(5), 987–991.

Rossi, P. E., & Allenby, G. M. (1993). A Bayesian approach to estimating household parameters. *Journal of Marketing Research, 30*(2), 171–182.
Rossi, P. E., Allenby, G. M., & McCulloch, R. E. (2005). *Bayesian statistics and marketing*. Hoboken NJ: Wiley (Wiley series in probability and statistics).
Selka, S., Baier, D., & Kurz, P. (2014). The validity of conjoint analysis: An investigation of commercial studies over time. In M. Spiliopoulou, L. Schmidt-Thieme, & R. Janning (Eds.), *Data analysis, machine learning and knowledge discovery* (pp. 227–234). Cham: Springer.
Steiner, W. (2010). A Stackelberg-Nash model for new product design. *OR Spectrum, 32*(1), 21–48.
Toubia, O., de Jong, M. G., Stieger, D., & Füller, J. (2012). Measuring consumer preferences using conjoint poker. *Marketing Science, 31*(1), 138–156.
Train, K. E. (2009). *Discrete choice methods with simulation* (2nd edition). Cambridge MA: Cambridge University Press.
Tsafarakis, S., Grigoroudis, E., & Matsatsinis, N. (2011). Consumer choice behaviour and new product development: An integrated market simulation approach. *The Journal of the Operational Research Society, 62*(7), 1253–1267.
Vermeulen, B., Goos, P., & Vandebroek, M. (2008). Models and optimal designs for conjoint choice experiments including a no-choice option. *International Journal of Research in Marketing, 25*(2), 94–103.
Voleti, S., Srinivasan, V., & Ghosh, P. (2017). An approach to improve the predictive power of choice-based conjoint analysis. *International Journal of Research in Marketing, 34*(2), 325–335.
Vriens, M., Wedel, M., & Wilms, T. (1996). Metric conjoint segmentation methods: A Monte Carlo comparison. *Journal of Marketing Research, 33*(1), 73–85.
Wang, X. J., Camm, J. D., & Curry, D. J. (2009). A branch-and-price approach to the share-of-choice product line design problem. *Management Science, 55*(10), 1718–1728.
Winkler, R. L., & Murphy, A. H. (1992). On seeking a best performance measure or a best forecasting method. *International Journal of Forecasting, 8*(1), 104–107.
Wirth, R. (2010). HB-CBC, HB-best-worst-CBC or no HB at all? In *Proceedings of the 15th Sawtooth Software Conference* (pp. 321–356).
Wittink, D. R., & Cattin, P. (1989). Commerical use of conjoint analysis: An update. *Journal of Marketing, 53*(3), 91–96.
Wittink, D. R., Vriens, M., & Burhenne, W. (1994). Commerial use of conjoint analysis in Europe: Results and critical reflections. *International Journal of Research in Marketing, 11*(1), 41–52.