Real World Performance of Choice-Based Conjoint Models

Martin Natter  
Markus Feurstein

Report No. 58  
June 2001
June 2001

SFB
‘Adaptive Information Systems and Modelling in Economics and Management Science’

Vienna University of Economics
and Business Administration
Augasse 2–6, 1090 Wien, Austria

in cooperation with
University of Vienna
Vienna University of Technology

http://www.wu-wien.ac.at/am

Papers published in this report series are preliminary versions of journal articles and not for quotations.

This paper was accepted for publication in:
European Journal of Operational Research

This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 (‘Adaptive Information Systems and Modelling in Economics and Management Science’).
Real World Performance of Choice-Based Conjoint Models

Abstract

Conjoint analysis is one of the most important tools to support product development, pricing and positioning decisions in management practice. For this purpose various models have been developed. It is widely accepted that models that take consumer heterogeneity into account, outperform aggregate models in terms of hold-out tasks. The aim of our study is to investigate empirically whether predictions of choice-based conjoint models which incorporate heterogeneity can successfully be generalized to a whole market. To date no studies exist that examine the real world performance of choice-based conjoint models by use of aggregate scanner panel data. Our analysis is based on four commercial choice-based conjoint pricing studies including a total of 43 stock keeping units (SKU) and the corresponding weekly scanning data for approximately two years. An aggregate model serves as a benchmark for the performance of two models that take heterogeneity into account, hierarchical Bayes and latent class. Our empirical analysis demonstrates that, in contrast to the performance using hold-out tasks, the real world performance of hierarchical Bayes and latent class is similar to the performance of the aggregate model. Our results indicate that heterogeneity cannot be generalized to a whole market and suggest that aggregate models are sufficient to predict market shares.

Keywords: Pricing Research, Choice Based Conjoint Analysis, Hierarchical Bayes, Latent Class, Heterogeneity

1 INTRODUCTION

Conjoint analysis is one of the most important tools to support product development, pricing and positioning decisions in marketing practice (cf. Wittink, Vriens and Burhenne, 1994; Wittink and Cattin, 1989). Researchers have developed different types of conjoint analysis (rating, ranking, choice based) as well as different techniques to estimate parameters of conjoint models. As compared to ranking or rating based conjoint approaches, choice based (Louviere and Woodsworth 1983) conjoint analysis (CBC) seems more realistic in imitating real shopping behavior. In CBC respondents have to choose one among several alternative products. This is, of course, a less difficult task for respondents and more like real shopping than rating or ranking alternatives. 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. However, CBC data are collected in interview situations, which may differ considerably from real shopping behavior. In a typical interview, respondents (i) are observed by an interviewer, (ii) simulate several purchases within a few minutes, (iii) are shown hypothetical assortments, and (iv) do not have any monetary consequences. Thus, estimates of CBC part-worths are not based on real purchase acts, but on simulated choices. Consequently, marketing actions that are based on such models may not lead to the same effects as in interviews. Furthermore, CBC-models are static whereas market shares may change over time. Dynamic effects and other impacts which are not captured within a CBC study, such as increasing brand awareness, changes in the level of distribution (cf. Golanty 1995), life cycle effects, promotional activity, seasonal impacts, new market entrants, etc., may decrease the real world validity of CBC models.

Practitioners typically use aggregate models to estimate CBC part worths and to build market simulators. However, it has been shown that models which consider heterogeneity can improve the internal predictive accuracy of CBC models (Huber 1998). The current marketing literature on choice modeling shows a strong focus on developing new CBC estimation methodologies such as Latent Class, Hierarchical Bayes or ICE (Individual Choice Estimation) (see, e.g., Johnson 1997, Hagerty 1985, DeSarbo, Ramaswamy and Cohen 1995, Lenk, DeSarbo, Green, and Young 1996, Arora, Allenby, and Ginter 1998, Pinnell 1994/1995). The major difference between these approaches lies in the extent to which they model heterogeneity of respondents. The internal validation of these new approaches has been tested in terms of hold-out tasks and Monte Carlo analysis (cf. Green, Krieger and Agrawal 1993, Vriens, Wedel and Wilm 1996, Garratt, Renken and Sigler 1998). The outcome of these studies prove the superiority of models which consider heterogeneity as compared to aggregate CBC models.
However, the real world performance of these new methodologies has not been examined so far. Consequently, we still do not know how good simulated purchases reflect real behavior. Neither performance on hold-out samples of the interview data, nor Monte Carlo analysis with artificial data (see, e.g., Vriens, Wedel and Wilms 1996) can help to answer this question. The research objective of our paper is the external validation of CBC models. In an empirical study, we compare real aggregate shopping data to predictions made by different types of CBC market simulators.

A paper related to our work is by Orme and Heft (1999) who investigate the external validation of three product categories. Developing market simulators based on aggregate logit, latent class and ICE (Individual Choice Estimation) models, they found that latent class and ICE predicted actual sales better than aggregate logit. Our study is different from Orme and Heft (1999) in two major aspects. First, we use real (national) aggregate data and not only the shops in which the interviews where made. While shop validation is interesting when a study is made for a single outlet, aggregate validation is more interesting for producing firms, product developers or retail chains. Second, Orme and Heft discarded 45 transitional weeks wherein prices had recently changed.

Kamakura and Ozer (2000) collect conjoint data from a sample of customers of a bank as well as actual banking behavior observed in the recent past from these same customers. These data allow them to directly compare estimates of preferences from various features of checking accounts with the same customers actual banking behavior. Using a multi-trait multi-method analysis, they test the external validity of part-worth estimates obtained with various estimation approaches. In contrast to the work of Kamakura and Ozer (2000) who compare estimates of preferences with actual behavior of the same group of customers, the aim of this paper is to compare shares of preferences obtained by different models to actual market shares. This is of major relevance to practitioners who conduct market surveys with a sample of the relevant customers and draw conclusions for the whole market based on this sample.

There is a prominent group of researchers (Carson et al. 1994; Neslin et al. 1994; Winer et al. 1994; DeSarbo, Ramaswamy and Cohen 1995; Orme and Heft 1999) who have indicated the need for additional research concerning external validity of conjoint analysis.

The key issue of our paper is the real world validation of two CBC estimation methodologies, latent class and hierarchical Bayes. Our empirical analysis is based on 43 different SKUs of four products.

The rest of the paper unfolds as follows: starting with the discussion of two alternative ways to estimate the model parameters (section 2), we proceed with the formulation of our hypotheses and measures (section 3). The description of our data (section 4) and results (section 5) is followed by an assessment of our outcomes.

2 MODELS

Researchers have proposed several techniques for CBC-parameter estimation. In this paper, we restrict ourselves to two of the more advanced methodologies, latent class (DeSarbo, Ramaswamy and Cohen 1995, Huber 1998) and hierarchical Bayes (Lenk, DeSarbo, Green, and Young 1996, Allenby, Arora, Ginter 1995, Allenby and Ginter 1995). For the purpose of comparability, we also report the results of an aggregate model (AG) as a special case of the latent class model.

These methods have been validated in terms of hold-out performance and with artificial data. Huber (1998), who compares hierarchical Bayes (HB) with latent class (LC) and extended latent class (ICE) finds that HB does best in terms of finding the correct parameters; accuracy of predicting hold-out choices was found to be similar for HB and ICE and slightly lower for LC models. Garrett, Renken and Sigler (1998) find that HB models are roughly four times as accurate in hit rate than logit. These results suggest that the incorporation of heterogeneity is an essential factor for improving consumer models. Moreover, it is known from Bayesian statistics that aggregate models underestimate the standard errors of the model parameters in the presence of heterogeneity. While HB models allow to model the individual consumer efficiently, it is not clear whether this methodology generalizes to a bigger world.

As long as one is interested in the behavior of an addressable individual (e.g., direct marketing),
the random coefficients (HB) based on the data can be used to predict her behavior in the future. When one wants to generalize to a larger group of individuals not from the sample, two alternatives are possible: When demographic information is available, one can try to assign individuals to the appropriate coefficients of the model. Rossi, McCulloch and Allenby (1996) included demographics in the hierarchy to determine if these could be used to explain individuals’ differences in behavior. However, their regression did not work very well. Subsequently, Ainslie and Rossi (1998) tried this with multiple categories, with not much more success. Without additional information of the consumers of interest, the random coefficients conditional on the data can not be used, but one has to take expectation values with respect to the distribution of heterogeneity. As compared to an aggregate model HB still provides different estimates. Suppose, we want to predict the share-of-preference of a specific SKU with attributes $x$. The functional form $f(x, \beta)$ of the choice probability with coefficients $\beta$ is supposed to be non-linear (MNL). In the case of a HB model, the share-of-preference is given by the expectation value of the choice probability taken over the distribution of heterogeneity ($E[f(x, \beta)]$). The coefficients of an aggregate model can be obtained from a random coefficient model by taking the expectation value of the random coefficients ($\beta = E[\beta]$) and the shares-of-preferences are $f(x, \beta)$. In order to see the differences between the aggregate and the random coefficient model one can expand the non-linear function $f$ in to a Taylor series around $\beta$: 

$$f(x, \beta) \approx f(\bar{\beta}) + (\beta - \bar{\beta})f'(x, \bar{\beta}) + 0.5(\beta - \bar{\beta})^2 f''(x, \bar{\beta}) \quad (1)$$

Taking expectation values, one obtains for the shares-of-preferences

$$E[f(x, \beta)] \approx f(x, \bar{\beta}) + 0.5\text{var}(\beta)f''(x, \bar{\beta}) \quad (2)$$

The second term $0.5\text{var}(\beta)f''(x, \bar{\beta})$ reflects the difference between an aggregate model and a random coefficient model averaged over the distribution of heterogeneity. The difference increases in the degree of non-linearity and the variance of the random coefficients (heterogeneity). Consequently, it can only be determined empirically whether models that consider heterogeneity perform better in a bigger world than aggregate models. This is the major focus of our analysis.

Latent Class

DeSarbo, Ramaswamy and Cohen (1995) propose to use a latent class version of CBC to overcome the limitations of aggregate analyses or a priori segmentations. The authors generalize the Kamakura, Russell (1989) scanner data response methodology to a latent class CBC model considering within subject replications over choice sets. Kamakura, Wedel and Agrawal (1994) introduced a latent class conjoint choice model which also included concomitant variables.

The respondent’s (segment specific) choice probability of SKU $j$ for segment $s$, $P_s(j)$, is given by

$$P_s(j \in C_n) = \frac{\exp(\beta_h(j, s) + p(j)\beta_p(s))}{\sum_{i \in C_n} \exp(\beta_h(i, s) + p(i)\beta_p(s))} \quad (3)$$

where $\beta_h(j, s)$ is the intrinsic utility of the SKU $j = 1, \ldots, J$ to segment $s = 1, \ldots, S$ and $\beta_p(s)$ the price utility for segment $s$. The “none-option or other SKU” has a price utility of zero.

$n = 1, \ldots, N$ choice sets
$C_n = \text{specific SKUs in the n'th choice set}$
$s = 1, \ldots, S$ market segments
$p(j)$ price of SKU $j$ in choice set $C_n$

For estimation of the CBC segment specific parameters we use the proposed maximum likelihood procedure. For the determination of the number of segments, we split the data into an estimation and a validation set (2 choice sets per respondent). The model with highest out-of-sample hit rate is chosen for real world validation (alternatively one could also use BAIC or other information criteria frequently applied in model selection).
Hierarchical Bayes

In contrast to the latent class model, individual-level estimates are obtained by a hierarchical Bayes model. Lenk, DeSarbo, Green, and Young (1996) showed analytically and empirically that hierarchical Bayes models do not require individual-level design matrices to be of full rank. This leads to the possibility of using fewer profiles per subject. Allenby and Lenk (1994 and 1995) applied hierarchical Bayes models to brand choice. The individual \((h)\) choice probabilities take the standard logit form:

\[
P_h(j \in C_n) = \frac{e^{\exp(\beta_h(j, h) + p(j)\beta_p(h))}}{\sum_{i \in C_n} e^{\exp(\beta_h(i, h) + p(i)\beta_p(h))}}
\] (4)

Heterogeneity across respondents is introduced by a multivariate normal distribution for the parameters (Arora, Allenby, and Ginter 1998) \(\beta_h = (\beta_h(., h), \beta_p(., h))\):

\[
\beta_h \sim \text{Normal}(\bar{\beta}, D)
\] (5)

The covariance matrix \(D\) is drawn from an inverted Wishart distribution. Using the Metropolis-Hastings algorithm the parameters are drawn from the posterior distribution:

\[
f(\beta_h|\bar{\beta}, D) \propto \exp\left(-\frac{1}{2}(\beta_h - \bar{\beta})' D^{-1}(\beta_h - \bar{\beta})\right) \Pi_n \Pr_{hn}
\] (6)

The probability that an individual \(h\) chooses choice set \(n\), is denoted by \(\Pr_{hn}\). In our study 1000 draws of the parameters separated by 10 iterations were taken after a thermalisation of 10000 iterations.

3 MEASURES AND HYPOTHESES

From the choice data, we build probabilistic choice simulators of the BTL-type and determine the shares-of-preference of all available products based on the real world prices of each period. The CBC estimates of the shares-of-preference based on market prices are then matched with the scanning data market shares. In order to test and compare the validity of choice-based conjoint approaches, we define one internal and three external criteria:

1. internal: \(hr\), the out-of-sample hit rate, which is an often used internal validation measure. This measure allows comparisons with other studies. Hit-rate is an interesting measure for management although it is based on the highest utility only.

2. external: \(r^2_{\text{ext}}\), the squared correlation between CBC-forecasts (Shares-of-Preference) and real scanning market shares (MS) when scanning prices are applied to the conjoint models. With pricing as the main purpose of these studies, only brand (SKU) and price utilities were estimated. Consequently, \(r^2_{\text{ext}}\) measures the quality of the price effects estimated from conjoint data.

3. external: \(\delta_{MS}\), the absolute deviation between average market shares and CBC Shares-of-Preference. \(\delta_{MS}\) measures the quality of the level estimates (brand (SKU) utilities). As several decisions in marketing and production planning depend on the forecasted market shares, this measure is especially interesting for new products.

4. external: In addition, we assess the external forecasting performance in terms of the Mean Squared Error (MSE) of the market share predictions compared to the real market shares. In contrast to \(\delta_{MS}\) the MSE measure stronger punishes larger deviations from the real market shares.

In accordance with the mentioned literature, we expect higher internal validity of hierarchical Bayes than latent class models. Furthermore, we hypothesize that both models (HB and LC) have a higher internal validity than an aggregate model (AG). Here, internal validity is measured in terms of out-of-sample hit rates, \(hr\). Hence, we predict that:
H1: HB estimation results in a higher out-of-sample hit rate than LC and aggregate estimation. Furthermore, we predict that LC has a higher out-of-sample hit rate than aggregate estimation.

Since most studies in this field only rely on internal validity measures, an interesting aspect of our analysis is the question whether internal validity measures are a good indicator for the real world validity of CBC models. If the correlation between internal and external validity measures is significantly different from zero, we can say that internal validity indicates higher real world validity. Accordingly, we hypothesize:

H2: There is a significant positive correlation between internal and external validation. Therefore, we expect that HB shows higher real world validity in terms of MSE and market share bias than the LC approach. Furthermore, we expect that the LC approach shows a higher real world validity than the aggregate approach.

Our discussion above (equation 2) also suggests that models that incorporate heterogeneity perform better.

4 DATA

In our analysis, we consider the following four product categories: mineral water (Cat1), shampoo (Cat2), shower gel (Cat3) and beer (Cat4). The conjoint-data analyzed here, were collected for four commercial pricing studies conducted between 1997 and 1999. The studies (see Table 1) differed in the number of (a) respondents (128/220/224/510), (b) choice sets per person (14/20/20/30) and (c) concepts per choice set (6/6/5/6). The stock keeping units (SKU) were characterized by brand, package size and price attributes. In all studies a none-option was included in the design. For each category, a randomized choice experiment that included scanned images of the products was programmed into Ci3, the Sawtooth Software questionnaire program.

The scanning data for these product groups with a total of 43 SKUs were provided by AC Nielsen. AC Nielsen collects data consisting of weekly sales and prices from a representative sample of stores and projects the sample data to nationally aggregated figures. The selection of representative samples and the weights used for the projection are only known to AC Nielsen. Therefore, we cannot test the representativeness of their sample here. However, since their customers can compare their own sales figures to the AC Nielsen reports, deviations are expected to be small. The scanning data are available for the same package types and package sizes as in the conjoint studies.

Table 1 shows the number of weeks (95/108/104/108), SKUs (11/13/10/9) and year and week of the last observation (97-25/98-22/98-50/99-12) for each product category. The majority of the SKUs is already established on the market and only 5 SKUs were introduced during the observed periods.

Since the CBC studies differed in the number of brands, package sizes and price levels, different encodings of prices and interactions were chosen. For the estimation of the conjoint models, the attributes were coded in the following way:

For Cat1, which included 2 package sizes (1 liter, 1.5 liter), we used a metric price variable for each package size, without any further consideration of price-brand interactions. The questionnaire for the shampoo data (Cat2), included brand specific price ranges which were determined from the real world price ranges. Therefore, we modeled price-brand interactions for Cat2. Shower gel (Cat3) is a category with many different package sizes. We used one metric price variable for each package size. In the beer study (Cat4), three different price ranges were built in the questionnaire and each SKU was assigned to one of three metric price variables. For categories 1, 3 and 4, we have also tested models with price-brand interactions. However, these models lead to similar results in terms of hit rate, so that we restrict the presentation of our results to the models without interactions for these three categories.
Table 1: Data set description

<table>
<thead>
<tr>
<th></th>
<th>Cat1</th>
<th>Cat2</th>
<th>Cat3</th>
<th>Cat4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mineral</td>
<td>shampoo</td>
<td>shower</td>
<td>beer</td>
</tr>
<tr>
<td></td>
<td>water</td>
<td>gel</td>
<td>gel</td>
<td>gel</td>
</tr>
<tr>
<td>SCANNING: nr. of weeks</td>
<td>95</td>
<td>108</td>
<td>104</td>
<td>108</td>
</tr>
<tr>
<td>SCANNING: nr. of SKU</td>
<td>11</td>
<td>13</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>SCANNING: year/week of last obs.</td>
<td>97/43</td>
<td>98/19</td>
<td>98/23</td>
<td>99/7</td>
</tr>
<tr>
<td>CBC: nr. of respondents</td>
<td>128</td>
<td>220</td>
<td>224</td>
<td>510</td>
</tr>
<tr>
<td>CBC: choice sets per person</td>
<td>14</td>
<td>30</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>CBC: concepts per choice set</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>CBC: year/week of last interview</td>
<td>97/25</td>
<td>98/22</td>
<td>98/50</td>
<td>99/12</td>
</tr>
</tbody>
</table>

5 CBC-VALIDATION RESULTS

Internal validity
Table 2 displays the overall and product category specific results (mean and standard deviation) of our study for the three methodologies (AG, HB, LC) and the four measures. The first column contains the product category followed by the internal (hit rate (out-of-sample)) performance measure and the real world fit measures ($r^2_{ext}$, $\delta_{MS}$, and MSE).

Our empirical results of the internal performance of the HB and LC estimation procedure is in line with the mentioned literature. Hierarchical Bayes estimation results in an out-of-sample hit rate of 65.0% whereas latent class$^1$ ends up with 38.1% only. This difference is significant (two-tailed t, 42 df) at the 1% level. As expected, the aggregate model - with a hit rate of 36.4% - performs worst. It’s performance is significantly lower than HB and slightly lower than LC. Therefore, based on our empirical results, hypothesis $H_1$ is confirmed. This advantage in hit rate may be due to simple shrinkage to individual level sample means (cf. Garratt, Renken and Sigler 1998).

Real world validity of the price utilities
Figure 1 plots SKU specific validity measures ($r^2_{ext}$) of HB versus those of LC. Points on the diagonal line show equal performance for both models. A result with a majority of points on one side of the diagonal would favor one of the methods. To our astonishment, HB and LC show the same real world performance in terms of the first external validity measure, $r^2_{ext}$. The t-statistic (t=-0.851, 42 df) shows that LC and HB do not have significantly different means$^2$. In contrast to the internal validation, the aggregate model ($r^2_{ext} = 49.6\%$) shows even slightly higher real world performance than the models that consider heterogeneity.

This result is reflected in the insignificant ($\alpha = 0.05$) correlation between the internal measure (hit rate) and the external measures ($r^2_{ext}$, $\delta_{MS}$). Consequently, $H_2$ is not supported empirically.

Real world validity of the brand utilities
The aggregate model shows a slightly lower performance in terms of market share level forecasts. However, with a $\delta_{MS}$ of 4.6% the aggregate model is not significantly different from HB and LC (see Table 2). Figure 2 plots the market share level errors of HB versus those of LC. Again, the diagonal line indicates indifference for either methodology.

The difference between average shares-of-preference and real average market shares, $\delta_{MS}$, is 4.5% for both, the LC models and the HB models. The t-statistic (t=0.144, 42 df) shows that the two approaches do not have significantly different market share level forecasts. We find a similar high correlation (r=0.91) between the HB and LC level errors as for $r^2_{ext}$. ANOVAs (see Tables 3, 4) with MSE and $\delta_{MS}$ as dependent variables and the estimation methodologies and the product categories as factors, confirm our findings, i.e., the choice of the methodology has no impact on the external predictive accuracy. However, the predictive accuracy varies significantly over the product categories. Thus, one cannot expect the same predictive accuracy for

1 For categories Cat1, Cat2, and Cat3, the highest hit rate was found for 3 classes. For Cat4, 2 latent classes where optimal in terms of hr.

2 They have a highly significant correlation of r=0.82.
different product categories. However, when the objective of a CBC market study is to build a market simulator, the choice of the methodology does not change the performance significantly.

The results show that the performance measures of the brands (SKUs) within a product category vary considerably. The correlation of 0.3 between the average market share of a brand and the external validity in terms of $r^2_{ext}$ indicates that small brands’ market shares are more difficult to predict than larger brands’ market shares. Further differences between brands within a category may be caused by dynamic effects (see above) which are not captured in a typical CBC study.

### 6 SUMMARY AND DISCUSSION

The aim of our study was to investigate empirically whether predictions of models which incorporate heterogeneity can successfully be generalized. Previous research has only shown that models like latent class or hierarchical Bayes clearly outperform aggregate models in terms of internal validity measures. Therefore, the question arises whether models which take heterogeneity into account, can still outperform aggregate models when applied to a bigger world (external validity). In our contribution, we showed theoretically that the answer is data dependent. Theoretically, the difference in performance increases with higher heterogeneity and higher non-linearity of the choice probability function (see equation 2). Therefore the question can only be answered empirically.

Up to date, this question has not been addressed directly by use of real world scanning data but by hold-out tasks only (Huber 1998). Our analysis is based on four commercial choice-based conjoint pricing studies including a total of 43 SKUs and weekly scanning data of the corresponding SKUs over approximately two years. An aggregate multinomial logit model served as a benchmark for two models that consider heterogeneity (latent class and hierarchical Bayes). The hit-rate for hold-out tasks was used as an internal performance measure. In accordance with previous research, our results underline the importance of heterogeneity: hierarchical Bayes ($hr=65\%$) clearly outperforms latent class ($hr=38\%$), which yields better forecasts than the aggregate model ($hr=36\%$). As a performance measure for the case that these models are applied to a whole market, we used the squared correlation ($r^2_{ext}$) between model predictions based on weekly market prices and weekly market shares from scanning data. As a measure for the level estimates, we calculated the absolute and squared deviation between average market shares and model forecasts. In contrast to the performance using hold-out tasks, the real world performance of hierarchical Bayes ($r^2_{ext} = 47\%$) and latent class ($r^2_{ext} = 49\%$) is similar to the aggregate model ($r^2_{ext} = 50\%$). The average bias of market share forecasts is 4.5\% for both, the Hierarchical Bayes and the latent class model and 4.6\% for the aggregate model. Our results indicate that heterogeneity cannot be generalized to a whole market. This is in contrast to the results of Orme and Heft (1999). The different outcomes may be explained by the data preprocessing steps performed by Orme and Heft. They retained only 71\% of the sales data where prices had not changed recently.

The CBC studies considered in this paper differ in several design factors (number of respondents, choice sets per person, concepts per choice set). Since in all of our 4 studies, the results between the different methods were comparable, we conclude that sampling deficiencies cannot explain why the aggregate performance of the HB and the LC model is not superior to a simple aggregate model.

One possible explanation for the difference in internal validity and market performance is that the hierarchical Bayes model better captures aspects specific in response behavior of subjects to the conjoint design, rather than the product itself.

Our empirical analysis indicates that at least one of the factors (heterogeneity or non-linearity in the choice probability function) which determine the difference between an aggregate and an individual level model is small.

In our contribution, the objective of the CBC market studies under consideration is to derive marketing implications for a whole market based on a representative sample of respondents. For marketing practice, our results suggest that aggregate models are sufficient to predict market shares.

In marketing applications where the consumers to be targeted are identical to the ones used to estimate the model parameters, latent class or hierarchical Bayes models are preferable. These

---

1In our four conjoint studies, for instance, the market does not allow for segment specific prices.
models can also help to identify potential market segments. However, once segments are identified, aggregate models for each segment are again sufficient to build a market simulator.

7 REFERENCES


Figure 1: The figure shows SKU specific external validity measures of Latent Class estimations versus Hierarchical Bayes estimations in terms of squared correlation between real world aggregate shop data and model predictions ($r_{ext}^2$). The diagonal line shows equal performance for both models. Since we cannot find a majority of points on one side of the diagonal, the result indicates comparable performance of both methods.
Figure 2: The figure shows SKU specific external validity measures of Latent Class estimations versus Hierarchical Bayes estimations in terms of market share bias ($\delta_{MS}$). The diagonal line shows equal performance for both models. Since we cannot find a majority of points on one side of the diagonal, the result indicates comparable performance of both methods.
Table 2: CBC-Validation Results: This table displays the overall and product category specific results (mean and standard deviation) of our study for the three methods (the aggregate model (AG), the Hierarchical Bayes model (HB) and the Latent Class model (LC)) and the four performance measures (the squared correlation between market shares and conjoint predictions ($r^2_{ext}$), market share bias ($\delta_{MS}$), the internal performance measure (hit rate (hr)) and the mean squared error (MSE)). The results clearly show that Hierarchical Bayes models outperform the Latent Class model and the aggregate model in terms of the internal measure (hr). However, the external measures do not show any significant differences between the models.

<table>
<thead>
<tr>
<th>Category</th>
<th>AG</th>
<th>HB</th>
<th>LC</th>
<th>AG</th>
<th>HB</th>
<th>LC</th>
<th>AG</th>
<th>HB</th>
<th>LC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hr</td>
<td>$r^2_{ext}$</td>
<td>$\delta_{MS}$</td>
<td>MSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>min.water</td>
<td>44.1</td>
<td>66.3</td>
<td>49.7</td>
<td>52.8</td>
<td>52.8</td>
<td>53.8</td>
<td>3.9</td>
<td>3.7</td>
<td>3.6</td>
</tr>
<tr>
<td>min.water</td>
<td>18.6</td>
<td>12.3</td>
<td>15.7</td>
<td>28.1</td>
<td>28.6</td>
<td>31.5</td>
<td>1.9</td>
<td>1.8</td>
<td>1.9</td>
</tr>
<tr>
<td>shampoo</td>
<td>6.1</td>
<td>65.9</td>
<td>4.4</td>
<td>53.1</td>
<td>49.3</td>
<td>46.6</td>
<td>3.9</td>
<td>3.7</td>
<td>4.0</td>
</tr>
<tr>
<td>shampoo</td>
<td>15.1</td>
<td>17.6</td>
<td>10.9</td>
<td>13.0</td>
<td>14.8</td>
<td>12.4</td>
<td>3.8</td>
<td>3.4</td>
<td>3.9</td>
</tr>
<tr>
<td>shower gel</td>
<td>25.7</td>
<td>54.7</td>
<td>56.1</td>
<td>49.3</td>
<td>49.1</td>
<td>50.6</td>
<td>6.1</td>
<td>6.1</td>
<td>5.9</td>
</tr>
<tr>
<td>shower gel</td>
<td>33.9</td>
<td>27.4</td>
<td>34.4</td>
<td>13.6</td>
<td>14.7</td>
<td>14.4</td>
<td>4.0</td>
<td>4.3</td>
<td>4.1</td>
</tr>
<tr>
<td>beer</td>
<td>52.7</td>
<td>73.5</td>
<td>52.6</td>
<td>41.2</td>
<td>34.8</td>
<td>42.8</td>
<td>4.8</td>
<td>5.2</td>
<td>5.0</td>
</tr>
<tr>
<td>beer</td>
<td>34.6</td>
<td>12.8</td>
<td>35.4</td>
<td>14.8</td>
<td>13.7</td>
<td>15.9</td>
<td>3.1</td>
<td>2.9</td>
<td>3.3</td>
</tr>
<tr>
<td>all</td>
<td>36.4</td>
<td>65.0</td>
<td>38.1</td>
<td>49.6</td>
<td>47.1</td>
<td>48.6</td>
<td>4.6</td>
<td>4.5</td>
<td>4.5</td>
</tr>
<tr>
<td>all</td>
<td>32.3</td>
<td>18.9</td>
<td>33.1</td>
<td>18.4</td>
<td>19.6</td>
<td>19.7</td>
<td>3.3</td>
<td>3.3</td>
<td>3.4</td>
</tr>
</tbody>
</table>

Table 3: This table shows the results of an ANOVA with the Mean Squared Error (MSE) as dependent variable and the estimation methodologies and the product categories as factors. The results show that the choice of the methodology has no impact on the external predictive accuracy.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
<th>Signif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1</td>
<td>79.760</td>
<td>0.012</td>
</tr>
<tr>
<td>Methodology</td>
<td>2</td>
<td>0.002</td>
<td>0.998</td>
</tr>
<tr>
<td>Product Category</td>
<td>3</td>
<td>5.093</td>
<td>0.002</td>
</tr>
<tr>
<td>Error</td>
<td>123</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: This table shows the results of an ANOVA with the market share bias ($\delta_{MS}$) as dependent variable and the estimation methodologies and the product categories as factors. The results show that the choice of the methodology has no impact on the external predictive accuracy.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
<th>Signif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1</td>
<td>260.737</td>
<td>0.000</td>
</tr>
<tr>
<td>Methodology</td>
<td>2</td>
<td>0.008</td>
<td>0.992</td>
</tr>
<tr>
<td>Product Category</td>
<td>3</td>
<td>3.589</td>
<td>0.016</td>
</tr>
<tr>
<td>Error</td>
<td>123</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>