MACRO CONSULTING, INC.
Marketing Research / Advanced Analytics

Abbreviated Task Sets:

Estimating Disaggregate Choice Models With Extremely Few Tasks Per Respondent

The author examines four commercial data sets to determine how few choice tasks and/or how few respondents are required for generating reasonably accurate disaggregate utilities when Hierarchical Bayes utility estimation is employed. The author demonstrates that, in carefully designed and analyzed studies, as few as one to four choice tasks per respondent can yield accurate disaggregate choice models.

Author: Paul Richard "Dick" McCullough

Published in the
CANADIAN JOURNAL OF MARKETING RESEARCH,
2003

Abbreviated Task Sets:

Estimating Disaggregate Choice Models With Extremely Few Tasks Per Respondent

Abstract

The author examines four commercial data sets to determine how few choice tasks and/or how few respondents are required for generating reasonably accurate disaggregate utilities when Hierarchical Bayes utility estimation is employed. The author demonstrates that, in carefully designed and analyzed studies, as few as one to four choice tasks per respondent can yield accurate disaggregate choice models.

Background

Conjoint analysis is a family of methods by which respondents' utilities for various product features (usually including price) are measured. In most cases, the utilities are measured indirectly. In these cases, respondents are asked to consider combinations of various product features, prices, brands, etc. as alternatives and state a likelihood of purchase or preference for each alternative (ratings- or rankings-based conjoint) or preference of one alternative over other competing alternatives (choice-based conjoint). As the respondent continues to make choices, a pattern begins to emerge which, through multivariate regression (and other) techniques, can be broken down and analyzed as to the individual features that contribute most to the purchase likelihood or preference. The importance or influence contributed by the component parts. i.e., product features, are measured in relative units called "utils" or "utility weights" or “partworths.”

In other cases, respondents are asked to tell the interviewer directly how important various product attributes are to them. For example, they might be asked to rate on a scale of 1 to 100 various product features, where 1 means not at all important to their purchase decision and 100 means extremely important to their purchase decision. The most popular direct measure is self-explicated scaling.

There are four main types of trade-off:

- Conjoint
- Discrete Choice
- Self-explicated
- Hybrid

This paper will focus on the application of Hierarchical Bayes (HB) estimation of individual-level choice-based conjoint models using three different choice methods: full-profile choice-based conjoint, full-profile constant sum choice and partial profile choice-based conjoint.

Hierarchical Bayes is a relatively new technique introduced into the academic literature in 1995 (Allenby, et al. 1995). The first commercial software application appeared in 1999 (Sawtooth Software 1998). Hierarchical Bayes "borrows" information from the total sample whenever an individual level model has insufficient information to be estimated independently.

Another technique with some applicability to choice-based conjoint disaggregate utility estimation is Latent Class Logit Regression (LC). LC attempts to capture heterogeneity by simultaneously segmenting the sample population based on their choice task responses and estimating segment-level logit model coefficients using maximum likelihood estimation. Andrews, et al. (2002) demonstrates that LC-based disaggregate choice utility estimation compares favorably to HB.

Choice-based conjoint is a form of conjoint where respondents are shown a set of alternative products and asked to pick one or none, just as they would if they were in a store. Each respondent is typically shown a series of choice tasks. Full-profile choice-based conjoint includes one level of each attribute in the study in each product alternative shown to respondents. Partial profile choice-based conjoint shows respondents alternatives that contain attribute levels from some subset of all attributes. In a constant sum choice study, respondents are asked to allocate a fixed number of points across all alternatives shown in each task.

Introduction

Conjoint Studies are renowned for yielding a great deal of strategic insight. In a single conjoint study, one can address price optimized relative to profit, revenues, unit sales or market share, optimal product feature set, cannibalization patterns, market response to competitor actions or new product introductions, dollar value of brand equity and many other issues.

However, conjoint studies capable of providing such rich findings are sometimes lengthy and/or time-consuming to field. A typical choice-based conjoint study, for example, will have 12-20 choice tasks (Sawtooth Software 1993) in addition to any other questions included in the survey. If the number of attributes is large, even more choice tasks may be desired. Often, conjoint exercises can be confusing and fatiguing experiences for respondents.

The development of Hierarchical Bayes techniques in the late 90s (Allenby, et al. 1995) not only has allowed the estimation of individual level utilities for choice-based conjoint but also for the more accurate individual level utility estimation of ratings-based conjoint. What has been generally ignored by the commercial research community is the fact that the efficiency of HB also allows for the reduction of the number of choice tasks required to support individual level utility estimation. Current practice is to design choice-based conjoint studies as if HB did not exist and then to apply HB to the resulting data. This is safe but, in some cases, as we will demonstrate, unnecessary.

Depending on the specific parameters of any given study, there are several potential advantages to reducing the number of choice tasks/conjoint ratings shown per respondent: 1) response rates may increase, 2) time in field may be shortened, 3) data collection costs may be reduced, 4) additional questions may be included in the survey to address other issues, 5) data may be collected in more modes, 6) Number of levels effect, order bias, learning bias, framing bias and respondent fatigue may all be affected by an abbreviated task set.

There are also several potential disadvantages. Clearly, the biggest potential disadvantage is that an inaccurate model may result. Further, Johnson and Orme (1996) demonstrate there are different response patterns in the early and late stages of a typical choice exercise sequence: 1) early responses tend to emphasize brand while later responses may emphasize price, 2) later responses may be better predictors than early responses, 3) respondent propensity to select "none" may increase in the later tasks.

However, if successfully executed, there are numerous potential applications for the abbreviated task set approach:

- Any study benefiting from a very brief interview but capable of generating a large sample:
  - Trade Show and Conference floor intercepts
  - Web surveys
  - Telephone surveys
  - Postcard surveys

- Studies combining conjoint with other issues such as segmentation, brand positioning or attitude and usage, resulting in an excessively long interview

- Realistic environment studies:
  - Laboratory simulations
  - Control store tests
  - Apparel fit tests
  - Taste Tests

The purpose of this paper is to demonstrate the feasibility of estimating disaggregate choice models using extremely few choice tasks per respondent.

Summary of Approach

Estimating a viable disaggregate choice model using from four to even one task per respondent is extremely ambitious. However, we will demonstrate, using four commercial data sets, that you can successfully reduce the number of choice tasks needed to estimate adequate disaggregate models if you follow the steps outlined below:

1. Generate a large sample size, most likely in excess of 500 or 1,000.
2. Carefully create an efficient experimental design:
   - Use as many versions as practical, that is, vary the choice tasks across respondents as much as you can.
   - Create numerous designs by varying the random seed in your experimental design software.
   - Evaluate these various experimental designs by comparing the design efficiency estimates by attribute. Pick the design with the largest minimum design efficiency.
   - Test your final design by creating a data set of random choices of the appropriate sample size. Estimate an aggregate choice model using this synthetic data. First confirm the model converges and then check the standard errors of each partworth.
3. Collect data.
4. Estimate utilities using HB:
   - Set burned iterations to 100,000.
   - Use monotonicity constraints whenever appropriate.

Research Method

For four commercial data sets, each of large sample size (n > 1,700), models were estimated using the total sample and all available tasks, excluding holdouts. Additional models were estimated for various reduced numbers of tasks and various smaller sample sizes. The sub-samples were generated by drawing independent random samples from the original data sets. Each resulting choice model was evaluated using Mean Absolute Error (MAE) and hit rates, where appropriate. All simulations were done using the share of preference method. Unless noted (Tables 9, 10 and 12), exponential tuning was not used on any of the models. No IIA bias correction factors were applied to any models. All four data sets included at least one holdout card.

Mean Absolute Error is the average of the absolute difference between each predicted and actual preference share for each holdout card (fixed task). When more than one holdout card is available, the reported MAE will be the average of the MAEs of all holdout cards. MAEs are calculated at the aggregate level.

MAEs are affected by the number of alternatives per task. In practice, assuming 3 to 5 alternatives per task, MAEs of 4 or 5 are typical and acceptable. With fewer alternatives per task, a slightly larger MAE would be acceptable. With more alternatives per task, the acceptable level would be slightly smaller.

Hit rates are the percentage of individuals for whom their preferred alternative in a holdout task is correctly predicted by their individual-level model. Hit rates are calculated at the individual level.

Hit rates are also affected by the number of alternatives per task. In practice, assuming 3 to 5 alternatives per task, hit rates of 60% or higher are typical and acceptable. With fewer alternatives per task, the acceptable level would be somewhat higher. With more alternatives per task, a slightly lower hit rate would be acceptable.

Each of the four data sets analyzed for this paper is described below in Table 1. The first three data sets, Beverages, Games and Books were initially designed with relatively few tasks while the fourth data set, Portals, has been designed with a more traditional number of tasks.

In the analysis, tasks were systematically excluded in such as way as to maximize the total number of unique tasks included in the data set in order to maximize design efficiency. In practice, this issue affected only the Beverages data set, which was conducted in-person with only three versions of the questionnaire. The other three studies were conducted online, allowing each respondent to receive one of 999 different choice task set versions.

To create the smaller samples used in this analysis, random draws were made from the total sample. Each task number/sample size combination was randomly drawn independently of the other data sets.

Table 1
Data Set Profiles

Data Set	Beverages	Games	Books	Portals
Sample Size	2,367	3,276	1,794	2,298
Attributes	15	18	5	6
Levels	57	42	16	24
Parameters	43	25	11	19
Random Tasks	4	6	4	12
Fixed Tasks	1	2	2	2
Alternatives per Task (excluding no-buy)	8	3	3	2
No buy alternative	no	yes	no	yes
Data Collection	In-person	Online	Online	Online
Survey Versions	3	999	999	999
Analytic Method	Constant Sum Choice	Partial Profile	Discrete Choice	Discrete Choice

Note that all HB runs were made with 100,000 iterations burned and every 10th of the next 10,000 saved. This number of burned iterations is much larger than typical. It was discovered early on in the analysis that HB does not converge as quickly when the number of tasks is reduced and that 100,000 burned iterations was a safe number to use for all data sets. This finding is reflected in Huber et al. (1998).

It should also be noted that computer run time was substantially lengthened by the increase in number of iterations. Run times for this analysis varied from two hours to 30 hours depending on study parameters and computer capabilities. Computers used in this analysis had clock speeds that ranged from 333 MHz to 1.2 GHz.

Also, monotonicity constraints were employed during utility estimation for three of the studies: Beverages, Books and Portals.

In all cases, the experimental designs were tested prior to field by estimating an aggregate multinomial logit (MNL) model using random data. Model convergence was confirmed and the standard errors of all partworths were examined for uniformity and magnitude.

Sample Size

Reducing the number of choice tasks will increase measurement error. Increasing sample size will decrease sample error. One hypothesis is that, by employing HB, we can use the increased precision from a larger sample size to counteract, at least to some degree, the decreased precision due to fewer choice tasks. Admittedly, there is substantial risk in sacrificing measurement precision for the sake of choice task parsimony. A critical issue examined in this paper is whether or not increased sample size can adequately compensate for this lost precision.

HB builds two models, an upper model, which is based on the total sample data, and a set of lower models, each one of which is based on a single person's data. HB "borrows" information from the upper model whenever the lower model does not fit the data particularly well. In this way, HB is able to estimate individual level choice utilities from data that is often too thin to estimate individual level utilities directly.

Table 2 below lists MAEs, for a study we recently conducted, Portals, calculated for a variety of sample size and number of tasks combinations. There appears to be an obvious trade-off between sample size and number of tasks. A sample size of 500, using 12 tasks, yields an MAE roughly equivalent to a sample size of 1,000 using 6 tasks, for example. At the very least, one can observe that MAE increases as sample size decreases and as number of tasks decreases.

The pattern observed below is consistent with the observation by Johnson and Orme (1996) "Apparently, one gets approximately the same increase in precision from increasing the number of tasks as from proportionately increasing the number of respondents." Although Johnson and Orme (1996) were encouraging the use of more tasks to increase precision, the data examined here suggest the reverse is also true. That is, increased sample size will also increase precision.

Portals

The Portals Study was an online conjoint study among registered users of an extremely large, well-known Internet portal. Registered users were sent an email inviting them to participate in an online study. For respondent convenience, a hyperlink to the online survey was embedded in the email invitation.

Respondents were shown 14 choice tasks. Each choice task contained two alternative portal configurations and a no-buy option. 12 of the choice tasks were used in the estimation of utilities and two were used as holdouts.

The MAEs in Table 2 below suggest that this study could have been conducted with as few as 3 random choice tasks (plus holdouts) if the sample were 2,298 and with as few as 4 random tasks if the sample were 1,000. All 12 random tasks were necessary to generate acceptable MAEs from a sample size of 500.

Table 2
Portals MAEs

n=	2,298	1,000	500	200
Tasks				
12	3.0	3.42	3.87	4.63
6	3.45	3.56	4.54	
4	3.96	4.3		
3	4.06	4.87		
2	5.49			
1	6.57			

In Table 3, notice that the hit rates drop gradually until the number of tasks is reduced to one. The hit rate of 1 task per respondent drops more sharply than at any other task reduction step.

Table 3
Portals Hit Rates

Tasks	Hit Rate (n=2,298)
12	86.1%
6	83.8%
4	82.1%
3	81.2%
2	79.9%
1	71.7%

For the original study design, the average standard error in the aggregate MNL model was 0.02 and the maximum was 0.03. As reference, the average and maximum standard errors for an aggregate MNL model with 3 tasks per respondent and sample size of 2,298 were 0.04 and 0.06, respectively. The average and maximum standard errors for an aggregate MNL model with 4 tasks per respondent and sample size of 1,000 are 0.06 and 0.07, respectively.

Experimental Design

When using an abbreviated task set, you need to take advantage of every opportunity to increase your model accuracy. Many experimental design software programs provide design diagnostics, typically an efficiency estimate. We used Sawtooth Software's CBC Advanced Design Module for designing all the studies reported here.

If you design a choice study with one set of tasks (one version) and compare the efficiency numbers from that design to another design using a large number of sets of tasks, say 999, you may find that the design efficiency is enhanced by the increased number of versions.

It is not always the case that an increased number of task set versions will increase design efficiency, of course, as some simple, small designs may achieve perfect balance and orthogonality with relatively few choice tasks. But in practice, most complex study designs can be noticeably improved by increasing the number of versions of choice task sets employed.

For that reason, when using the abbreviated task set approach outlined here, use as many questionnaire versions (number of task sets) as is practical.

Even with a complex study design, a large number of versions of the choice tasks is not necessarily a critical factor, however. We recently designed a study, Beverages, with only three survey versions. It was a paper and pencil study with expensive visual exhibits which limited the number of survey versions we could employ. We used four random tasks, estimated 43 parameters (57 total attribute levels) and had an MAE of 1.05 (see Table 6 below).

When designing your choice tasks, most experimental design software uses a randomized search algorithm. The software creates thousands of possible choice tasks and efficiently searches through them, selecting and then discarding tasks until an adequate set is found. When you have a large number of tasks, and if your experimental design is straightforward, whatever the software

generates is probably going to be adequate. But if you're using a limited number of choice tasks, like we are here, we suggest changing the random seed your software starts with 15 or 20 times. You'll get a different design with each new random seed. The design efficiency of those 15-20 designs may vary noticeably. Select the design with the best minimal partworth efficiency.

Design efficiency is not sufficient to guarantee that your experimental design will yield an adequate choice model, however. For example, if you create a design with partworth efficiencies all equal to 1 but you have an extremely small sample size, you may not be able to estimate your logit model. Design efficiency calculations have nothing to do with sample size but sample size has a great deal to do with whether or not a model is estimable.

One simple way to determine if your model will successfully generate an adequate disaggregate choice model is to test it using synthetic data. To do this, create a data set of random choice data. Make as many records as you expect to have sample. Estimate an aggregate MNL model. First, check to see if the model converges. Assuming it converges, check the standard errors of each partworth. Our experience suggests that if all the partworth standard errors are 0.05 or less, a good model will result when the real data comes in.

Games

The Games Study was an online conjoint study among registered users of a particular online gaming site. Registered users were sent an email inviting them to participate in an online study. For respondent convenience, a hyperlink to the online survey was embedded in the email invitation.

Respondents were shown eight choice tasks. Each task contained three alternative gaming sites and a no-buy option. Six choice tasks were used in utility estimation and two tasks were used as holdouts.

The Games study was designed using partial profile choice. Approximately one-third of the total number of attributes were represented at any one time. Thus, a sample of 1,000 in the Games data set is roughly equivalent to a sample size of 330 using a full-profile data set, in terms of attribute level exposure. The robustness of HB is evident in its ability to estimate good models with sample sizes as low as 200 and as few as 3 tasks (Table 4) for this partial profile design.

Note that where larger sample size or greater number of tasks yields MAEs of 4 or greater, further MAEs are not calculated. Also note there is some slight instability in MAE estimates due to sampling error in the subsample draws.

Table 4
Games MAEs

n=	3,276	2,000	1,000	500	200
Tasks					
6	1.79	1.93	1.87	2.41	2.56
4	1.85	2.51	2.83	3.33	3.59
3	2.63	2.64	2.75	3.4	3.85
2	3.52	3.36	3.34	3.61	5.37
1	4.47	5.89	4.94	14.19	

Hit rates are unusually high (see Table 5). This is most likely due to a large no-buy share. What is noteworthy, however, is the modest decline in hit rate as task number decreases.

Table 5
Games Hit Rates

Tasks	Hit Rate (n=2,298)
6	81.4%
4	79.9%
3	78.4%
2	78.3%
1	76.3%

This study was originally designed for 6 tasks per respondent, with 999 versions of the choice task set, yielding virtually unique questionnaires per respondent. Sample size was expected to approach 2,000. By examining the standard errors of the partworths in the aggregate MNL model estimated with synthetic, i.e., random, data of the size n=2,000, the design was expected to provide an accurate model with 6 tasks per respondent.

For the original study design (6 tasks and 2,000 respondents), the average standard error in the aggregate MNL model was 0.02 and the maximum was 0.03. As reference, the average and maximum standard errors for an aggregate MNL model with 1 task per respondent and sample size of 3,276 were 0.04 and 0.06, respectively. The average and maximum standard errors for an aggregate MNL model with 2 tasks per respondent and sample size of 200 were 0.13 and 0.16, respectively.

Beverages

The Beverages Study was conducted among grocery shoppers in a South American country. Respondents were shown a series of 5 boards depicting 8 different beverage products they might buy in a grocery store. Four of the boards were used to estimate individual level choice utilities. The fifth board was used as a holdout task.

The interviews were personal, one-on-one interviews conducted in six regions within the South American country. Sample size was approximately 400 per region. Respondents were shown a board of 8 alternative beverages and asked how many of each they would buy if these were the beverages available to them in the grocery store they typically frequented. This numeric data was converted to constant sum for the purpose of utility estimation.

One unusual aspect of this study was the high quality of exhibits shown to respondents. Products were depicted visually. Most study attributes did not need to be listed explicitly on a card but were incorporated into the visual shown. For example, one attribute was container structure. Rather than listing the word “can” or “bottle” on a card and asking preference, respondents were shown some products in cans and others in bottles.

For the Beverage study in Table 6 below, MAEs of under 3 were obtained using just 1 choice task per person with a sample size of 2,367 or 2 tasks per person with a sample size of 500. This is particularly remarkable considering the large number of parameters to be estimated (43) and the small number of questionnaire versions available (3).

Table 6
Beverages MAEs

n=	2,367	1,000	500	200
Tasks				
4	1.05	2.11	2.42	3.94
2	1.86	2.2	2.9	3.75
1	2.89	4.33	6.28	10.29

This study was originally designed for 4 tasks per respondent, with three versions of the choice task set, yielding 12 different choice tasks. Sample size was expected to approach 2,400. By examining the standard errors of the partworths in the aggregate MNL model estimated with synthetic, i.e., random, data of 2,400 cases, the design was expected to provide an accurate model with 4 tasks per respondent.

The average standard error in the aggregate MNL model was 0.06 and the maximum was 0.11. As reference, the average and maximum standard errors for an aggregate MNL model with 2 tasks per respondent and sample size of 200 were 1.3 and 6.9, respectively.

Note that to construct the aggregate MNL model, while it was not necessary to assume the Beverages Study design was that of a full-profile choice-based conjoint, rather than constant sum, we did so for reasons of convenience. It could be assumed the true standard errors are somewhat smaller since constant sum choice provides more information per task than discrete choice.

It is possible that the large number of alternatives per task, the constant sum choice analytic method and/or the visual expression of choice tasks contributed to the success of the study.

Hit rates were not calculated for the Beverages data set because hit rates are not appropriate for constant sum data.

Estimate Utilities

Books

The Books Study was an online conjoint study. Respondents were shoppers of a particular bookstore. Shoppers were sent an email inviting them to participate in an online study. For respondent convenience, a hyperlink to the online survey was embedded in the email invitation.

Respondents were shown six choice tasks. Each choice task contained three alternative bookstores. There was no no-buy option. Four of the choice tasks were used in the estimation of individual-level utilities and the remaining two choice tasks were used as holdouts.

The Books model, as shown in Table 7, has the poorest untuned MAEs of any data set examined. However, with only 4 tasks per person and given a fairly large sample size of 1,794, the MAE of 4.32 is marginally acceptable. Note that the MAE estimates at smaller sample sizes n=500 and

n=200 were extremely volatile and not reported. Asterisks were inserted to denote instability. This is most probably due to a combination of sampling error and relatively poor model performance.

Table 7
Books MAEs

n=	1,794	1,000	500	200
Tasks				
4	4.32	5.03	**	**
3	5.66	6.45	**	**
2	7.96		**	**

Hit rates were again extremely high, most likely due to the dominance of one brand in the holdout tasks. However, notice the very modest declines in hit rates as number of tasks decreases.

Table 8
Books Hit Rates

Tasks	Hit Rate (n=1,794)
4	87.7%
3	87.04%
2	86.73%

As was the case with the other studies, the Books design was tested prior to field by estimating its MNL model using synthetic data.

For the original study design (4 tasks and 1,794 respondents), the average standard error in the aggregate MNL model was 0.02 and the maximum was 0.03. As reference, the average and maximum standard errors for an aggregate MNL model with 3 tasks per respondent and sample size of 1,794 were 0.03 and 0.04, respectively. The average and maximum standard errors for an aggregate MNL model with 4 tasks per respondent and sample size of 1,000 were 0.03 and 0.04, respectively.

The relatively large MAE at 4 tasks per respondent may be due to the small number of attributes in the study failing to model respondents' choice behavior and/or the failure to include the most relevant attributes to respondent choice behavior in the study. A qualitative examination of the attributes would suggest the latter alternative as a potential explanation for the relatively large MAE value.

A third possible explanation for the relatively poor performance of the Books model is that there may have been excessive heterogeneity in the data set which, given a small number of choice tasks, would have put undue pressure on HB's upper model. The lower model could not estimate accurate utilities because there were too few tasks. The upper model could not act as an adequate surrogate because of sample heterogeneity.

Sentis and Li (2001) examined seven different commercial data sets and found that segmenting the data, via Latent Class, prior to HB did not improve the HB results for any of the seven data sets. However, the commercial studies they examined were designed with a more orthodox

number of tasks than the studies examined here. When using an extremely small number of choice tasks, it appears plausible that the LC/HB combination might indeed provide better utility estimates than HB alone.

To test this third hypothesis, we went back to the Books data and, using the 4 random tasks as basis, identified a four-segment Latent Class solution. We reran HB within each segment and pooled the resulting utilities. Similarly for Portals, using only 2 random tasks, we created a two-segment Latent Class solution, reran HB within segment and pooled the resulting utilities.

For both Latent Class solutions, we used monotonicity constraints, evaluated one through 10 segment solutions, replicated each solution five times and set the iterations limit to 250. In both cases, we selected the number of segments solution with minimum CAIC.

To insure valid comparability across models, each model in Tables 9, 10 and 12 were exponentially tuned to minimize average MAE.

It is clear from the data in Table 9 below that, for these specific studies, the combination of LC and HB did not generate better share preference estimates than did HB alone. Further, hit rates in both cases were slightly poorer for the LC/HB approach than for HB alone.

Table 9

Study	Tasks Used In	Tasks Used In	Error Term	NB within LC	Tuning Scalar	HB Alone	Tuning Scalar
Books	4	4	MAE Hit Rate	1.36 85.5%	0.48	1.14 87.7%	0.54
Portals	2	2	MAE Hit Rate	1.53 76.9%	0.50	1.44 79.9%	0.34

Perhaps LC/HB did not improve MAEs in these studies because there were a very small number of choice tasks on which to base any segmentation. It still seems possible, then, in the context of extremely few choice tasks per respondent, that HB run within highly homogeneous segments may outperform HB alone.

To test this hypothesis, we went back to the Portals study and, using all 12 original random tasks as basis, identified a six-segment Latent Class solution. We reran HB within each segment and pooled the resulting utilities. Table 10 shows the results of this new analysis compared to the data from Table 9.

Table 10

Study	Tasks Used In	Tasks Used In	Error Term	NB within LC	Tuning Scalar	HB Alone	Tuning Scalar
Portals	2	2	MAE Hit Rate	1.53 76.9%	0.50	1.44 79.9%	0.34
Portals	2	12	MAE Hit Rate	1.04 80.2%	0.29	1.44 79.9%	0.34

The LC segmentation based on 12 tasks generated superior MAEs to both the original, unsegmented model and also the 2 task segmented model. Further, hit rates, while only slightly improved relative to the original estimates, are substantially better than the 2-task model. This seems to suggest that increased sample homogeneity can potentially improve HB estimates of both MAEs and hit rates.

These improvements in both MAEs and hit rates were achieved despite the decrease in sample size that resulted from using a six-segment solution rather than a two segment or no segment solution. This seems to add further credence to the hypothesis that sample homogeneity can potentially improve HB estimates when few choice tasks per respondent are employed.

Because HB is borrowing heavily from the upper model, more iterations are required to gain convergence. So, to be safe, set your burned iterations to 100,000, as did Huber, et al. (1998). When you use an abbreviated task set, you are not likely to get convergence as quickly as you would with a more conventional number of tasks.

As an aside, this result may shed some light on Sentis and Li (2000). In that paper, they observe that you only need burn as few as 1,000 HB iterations, not 20,000 iterations, as is generally assumed. We suspect that, because typical commercial data sets use a larger number of choice tasks than is minimally necessary, HB did not need to borrow as much from the upper model and, therefore, converged faster.

We subjectively reviewed 69 log files from HB runs made for this paper. Sawtooth provides four measures of fit in its HB log files. Based on years of experience, we've found that that the first two diagnostics, percent certainty and root likelihood seem to converge quicker than the last two, average variance and parameter RMS. So to simplify the task and yet remain conservative, we scanned the parameter RMS statistics from all 69 log files to qualitatively determine when convergence was achieved. We then regressed the number of iterations at which convergence was subjectively determined to occur against sample size and number of tasks.

Table 11 clearly indicates that sample size does not appear to affect how many iterations are needed for convergence but the number of tasks has a substantial impact on the number of runs needed for convergence. This finding is entirely consistent with the hypothesis that more iterations are needed to obtain convergence when the lower model has less individual-level information available.

Table 11

Variable	B	Std Error	Beta	t	Sig
Sample Size	-0.004	0.004	-0.120	-1.070	0.288
Number of Tasks	-9.252	2.283	-0.456	-4.052	0.000

Of the 69 HB logs examined, the average convergence was at roughly 60,000 iterations. The range was from 4,000 to 100,000.

Although we have not verified its impact, we suggest you use monotonicity constraints when estimating your utilities. Orme and Johnson (1997) review numerous studies that indicate the use of constraints "can significantly improve the predictive ability of full-profile conjoint utilities."

Johnson (2000a) elaborates, "Constraints usually provide improvements in hit rates and sometimes provide improvements in share predictions." Additionally, the absence of monotonicity constraints may undermine the confidence of the end-user of the study results if unconstrained partworths have incorrect slopes.

However, if the researcher is primarily concerned with share predictions, is relatively unconcerned with predicting individual choices and is also unconcerned with end-user confidence, he/she may want to consider not including monotonicity constraints because they do not always improve share predictions and may, in some cases, harm share predictions.

All the data reported here, except for that from the Games study, were based on estimations using constraints, both in the HB runs and the Latent Class segmentations.

As a final point, the reflective reader might ask what's the difference between reducing the number of tasks to one to estimate a disaggregate model and simply running an aggregate model. The answer is a disaggregate model estimated with an abbreviated task set may still outperform the aggregate model. Simply put, even with a small number of choice tasks, we may still be capturing heterogeneity with the disaggregate model that we may not be capturing with the aggregate model.

Table 12

			Aggregate Model			Disaggregate Model	
Study	**n**	**MAE**	**Tuning Scalar**	**Tasks**	**MAE**	**Tuning Scalar**	**Tasks**
Books	1,794	1.68	1.2	4	1.14	0.54	4
Books	2,298	0.87	0.94	12	0.52	0.38	4

Discussion

This paper has examined the practicality of conducting commercial choice-based conjoint studies with extremely few tasks per respondent by manipulating sample size and number of tasks. The robustness of HB is evident in the quality of the results examined. In all of the data sets presented here, relatively accurate individual-level choice models were created using no more than four tasks per respondent, given sample size of 1,000 or more.

There are other factors, in addition to sample size and number of tasks, that may affect model error:

- Large variety of tasks (number of task set versions)
- Aggregate Customization experimental design technique (Arora and Huber, 2001)
- Other trade-off techniques, such as, max-diff or extent-of-preference

Which, if any, of these factors significantly affect model performance?

Nothing is known about the impact of the underlying distributions on the technique outlined here, either. Further research needs to be done on more commercial data sets as well as synthetic data sets to better understand the impact of these issues.

Results of this study may offer additional hypotheses concerning two findings recently published:

- Sentis and Li (2000) reported HB convergence for several commercial data sets after as few as one thousand iterations
- Sentis and Li (2001) reported that, again for numerous commercial data sets, HB alone performed as well as Latent Class followed by HB within LC segment

In both cases, these findings may be the result of the number of choice tasks used. HB may converge more quickly when there is an abundance of individual-level data. It appears clear that the reverse is also true, namely, that when fewer tasks are used, a larger number of iterations is required to reach convergence.

Similarly, Latent Class segmentation may not offer much assistance in those cases where the individual-level model is information rich. That is, where the upper level HB model does not contribute much to the lower level model.

Conversely, increasing sample homogeneity by segmentation prior to running HB does appear to potentially improve model performance, when very few tasks per respondent are employed and an adequate segmentation can be created. However, in practice, developing an optimal segmentation scheme given very few choice tasks per respondent appears somewhat problematic. If a large number of choice tasks per respondent are used, LC segmentation does not appear to be helpful. If very few tasks are used, there is little data on which to base the segmentation. Perhaps in the future other methods may be developed for constructing homogeneous segments that don't rely exclusively on choice tasks. If so, the potential for the LC/HB approach may be substantially enhanced.

A related question is how well Latent Class alone estimates individual-level utilities, relative to HB, given an abbreviated task set and large sample size. Andrews, et al. (2001) demonstrated that Latent Class may out-perform HB under certain conditions.

The Beverages data set performed particularly well. The Beverages study differed from the other studies in several ways:

- Constant sum choice
- Large number of alternatives per task
- In-person interview
- Visual representation of products (rather than written descriptions)
- Respondents were South American (all three other studies were conducted in the USA)

Which, if any, of these factors significantly contributed to the unusually strong performance of the Beverages model?

Several biases thought to be inherent in conjoint studies, namely number of level effect, order bias, learning bias, framing bias and respondent fatigue, may all be affected by an abbreviated task set. Further study needs to be undertaken to determine whether or not and if so, to what degree, any of these biases might be affected by the use of abbreviated task sets.

Summary

It appears that surprisingly good individual-level choice models can be constructed with as few as one to four choice tasks per respondent when using HB.

This approach requires a careful effort, involving these factors:

- Large sample size, in excess of 500 or perhaps 1,000.
- Highly efficient experimental design.
- As many survey versions as is practical.
- Design tested with random choice data for the same sample size as expected from the field (keep standard errors under 0.05, more or less).
- Run HB
- Large number of burned HB iterations, perhaps 100,000.
- Monotonicity constraints.

Computer run times can be significantly and adversely affected by the increase in sample size and burned iterations.

There are numerous practical situations where sample size is more easily attainable than a large number of choice tasks per respondent. In those situations, the reduced task set approach may prove valuable and useful.

References

Allenby, G. M., Arora, N., and Ginter, J. L. (1995) "Incorporating Prior Knowledge into the Analysis of Conjoint Studies," Journal of Marketing Research (May), American Marketing Association, Chicago, IL.

Andrews, Rick L., Ansari, Asim, and Currim, Imran S. (2002) "Hierarchical Bayes Versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery," Journal of Marketing Research (February), American Marketing Association, Chicago, IL.

Arora, Neeraj, and Huber, Joel (2001) "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," Journal of Consumer Research (September).

Huber, J., Arora, N., and Johnson, R. (1998) "Capturing Heterogeneity in Consumer Choices," 1998 Advanced Research Techniques Forum Proceedings, American Marketing Association, Chicago, IL.

Johnson, Richard M. (2000a) "Monotonicity Constraints in Choice-Based Conjoint with Hierarchical Bayes," Sawtooth Software, Inc., Sequim, WA.

Johnson, Richard M. (2000b) "Understanding HB: An Intuitive Approach," 2000 Sawtooth Software Conference Proceedings, Sawtooth Software, Inc., Sequim, WA.

Johnson, Richard M. (1999) "The Joys and Sorrows of Implementing HB Methods for Conjoint Analysis," Sawtooth Software, Inc., Sequim, WA.

Johnson, Richard M. and Bryan K. Orme (1996) "How Many Questions Should You Ask In Choice-Based Conjoint Studies?" 1996 Advanced Research Techniques Forum Proceedings, American Marketing Association, Chicago, IL.

McCullough, Dick (2003) “Getting the Most Bang from the Fewest Questions: A New Approach to Designing and Analyzing Conjoint Studies,” 003 Professional Marketing Research Society Conference Proceedings, Vancouver, B.C.

McCullough, Dick (2002) “A User’s Guide To Conjoint Analysis”, Marketing Research (Summer), American Marketing Association, Chicago, IL.

McCullough, Dick (2001) “Trade-off Study Sample Size: How Low Can We Go?,” 2001 Sawtooth Software Conference Proceedings, Sawtooth Software, Inc., Sequim, WA.

Orme, Bryan (1998) “Sample Size Issues for Conjoint Analysis Studies,” Sawtooth Software, Inc., Sequim, WA.

Orme, Bryan and Richard M. Johnson (1997) “Using Utility Constraints to Improve the Predictability of Conjoint Analysis,” Sawtooth Software, Inc., Sequim, WA.

Sawtooth Software (1993) “The CBC User Manual,” Sawtooth Software, Inc., Sequim, WA.

Sawtooth Software (1998) “The CBC/HB Module for Hierarchical Bayes Estimation,” Technical Paper accessible from sawtoothsoftware.com web site.

Sentis, Keith and Lihua Li (2000) “HB Plugging and Chugging: How Much Is Enough?” 2000 Sawtooth Software Conference Proceedings, Sawtooth Software, Sequim, WA.

Sentis, Keith and Lihua Li (2001) “One Size Fits All or Custom Tailored: Which HB Fits Better?” 2001 Sawtooth Software Conference Proceedings, Sawtooth Software, Sequim, WA.

Published in the CANADIAN JOURNAL OF MARKETING RESEARCH, 2003

MACRO
CONSULTING, INC.

We are an independent ***marketing research consulting firm*** *dedicated to helping you make the most informed, insightful marketing decisions possible. We specialize in technology, consumer, and new product research, and are well recognized for our* ***State-of-the-Art Research*** *techniques.*

Ultimately, we provide more than just technical expertise. We focus on developing ***pragmatic solutions*** *that will have a positive impact on the profitability of our clients.*

CONTACT US:

Telephone: 650-823-3042

General Inquiries:
info@macroinc.com

Advanced Analysis Inquiries:
analysis@macroinc.com

richard@macroinc.com

www.macroinc.com