Sawtooth Software

RESEARCH PAPER SERIES

History of Sawtooth Software's CBC Program

Rich Johnson,
Sawtooth Software, Inc.

530 W. Fir St.
Sequim, WA 98382
(360) 681-2300
www.sawtoothsoftware.com

History of Sawtooth Software's CBC Program

Rich Johnson, Chairman
Sawtooth Software, Inc.

Before Sawtooth Software, I had worked for about 20 years in the market research industry, mostly developing new methods of analysis. In the early 1980s I retired from that life and, with my wife, Judy, moved from Chicago to Sun Valley, Idaho. In 1983 we formed Sawtooth Software, named for the nearby Sawtooth Mountains, and I began work on what would become our first product, the Ci2 System for Computer Interviewing.

During the next few years I was joined by several former colleagues, and together we introduced many new software products. My previous experience with actual client problems had given me a sense of what kinds of software would be useful to market researchers. But, after nearly a decade of intensive focus on software development, I felt the need to become reacquainted with current real-world problems. In the early 1990s I stepped down from management of Sawtooth Software and began consulting with its users, hoping to gain experience about current problems that might inspire ideas for new products.

The Beginnings of CBC

One software user with whom I consulted was John Fiedler. Over the years, John had proven to be an imaginative and innovative researcher, and was an ideal partner with whom to explore new territory. John had developed a relationship with an aircraft manufacturer who had a pressing problem. That problem provided the opportunity to develop a prototype for what would eventually become CBC.

The legal environment in the U.S. had encouraged lawsuits against aircraft manufacturers, even for accidents which were clearly due to pilot error. In response, most manufacturers of propeller-driven aircraft for the general aviation market in the U.S. had ceased production several years earlier. Now, however, the legal environment had changed in a way that could benefit aircraft manufacturers, and this manufacturer was examining the possibility of re-entering the market with the reintroduction of a formerly popular model.

Our potential client needed to forecast the number of planes likely to be purchased per year at each of several possible prices. There was considerable uncertainty about how to price this plane. Several years of inflation had occurred in the interim, and it seemed likely that potential buyers might regard a correspondingly higher price as unreasonable. We considered various approaches that might be used to answer this question. ACA had achieved wide popularity by that time, but pricing research had not been one of its strong points. We considered asking respondents directly what they would pay for such a plane, but rejected that approach because I had believed for many years that the best pricing results are obtained when the respondent is not aware that pricing is the object of the study.

I had been aware of Jordan Louviere's work with choice experiments, and the basic idea of having respondents choose among sets of concepts with different features at different prices seemed appealing. It was a task that eager airplane buyers might find interesting, and that approach could put price in a context that mirrored actually having to pay for the airplane. Additionally, I had acquired previous experience with choice-based conjoint. About ten years earlier, I had been involved in several studies in the beverage industry to investigate effects on purchase likelihood of brand, price, and package size. In those studies we had respondents choose among sets of product concepts, and we found the resulting data to be useful in forecasting market response to pricing changes.

John and I decided to propose such a study for the aircraft manufacturer. This was, of course, long before the Internet, when an efficient way to do interviews was "disk by mail," in which the researcher mailed floppy disks to respondents who inserted the disks in their computers, took the interview, and mailed the disks back. The client's director of marketing research briefed us on the various features being considered for the plane in question, as well as the range of prices contemplated. We programmed a sample interview using Ci2, which asked the respondent to make choices among several sets of concepts described by levels of those attributes. We used Ci2's ability to randomize levels of concepts, so that every respondent had a unique questionnaire, the attributes varied independently, and within each attribute, every level appeared the same number of times. When we were satisfied with that questionnaire, we arranged a meeting to present our proposal to the President of the company.

In the initial stages of the meeting we sensed quite a lot of skepticism. But we persuaded the President to take the interview, as though he were in charge of purchasing for a company who operated a fleet of similar but aging aircraft. He warmed to the task. In the end he saw that the interview did mimic the actual purchase process, and was persuaded to undertake a study with real respondents.

The sample consisted of owners of the manufacturer's previously produced small aircraft, plus "fixed base operators" throughout the U.S., who serviced small aircraft. Data collection went well. In that industry many of the players knew one another, and the disks went out with personal letters from the President asking for cooperation. The response rate was good and the data seemed to be of high quality.

For the presentation of our results, we built a simulator to estimate the expected number of sales per year for a plane with any combination of features and prices. The data indicated quite clearly which features should be included in the product, and those made sense to the client. However, our sales predictions were much lower than desired. The client had a sales target in mind, and we found that to sell that many planes would require prices lower than planned. In the client's mind, the project was probably only a limited success. It picked features well, but suggested a price which would generate less revenue than desired.

The client decided to ignore the pricing information from the study, and to market the plane at the higher price originally planned. Later, we learned informally that there had been production problems which limited the number of planes available. That reduced supply of planes did sell at the higher price. If asked how to price planes so as to sell only that smaller number, our model would have chosen a price not much different from that which was actually used. Thus, the validity of our projections seemed to be confirmed, though only informally.

Our experience with the aircraft study was very positive. Respondents seemed to enjoy the choice tasks, data were gathered quickly, the response rate was high, analysis proceeded without a hitch, and we were able to report useful results to our client. The client was disappointed with some of the findings, but events later confirmed that the findings were probably correct. In addition to the satisfaction of serving a client, I had the benefit of discovering the very thing I had been searching for – the germ of an idea for a new software product. I began to think about what characteristics such a product should have, and even began to write code, but did not retreat entirely into the world of programming. I continued to work with other Sawtooth Software users, and had many opportunities to use a CBC-like approach for their clients. Thus, when eventually introduced, CBC had already been tested in a number of projects.

Designing a New Software Product

At this point the design of CBC became a group enterprise, benefiting from knowledge of others within Sawtooth Software. Many of our software users were generalists rather than experts in statistical analysis. Accordingly, we thought this product, like our others, should be made easy to use and hard to misuse. We thought it would be desirable to handle questionnaire design automatically. Our goal was a system in which the user would specify the attributes and levels, as well as the number of choice tasks and the number of concepts to appear in each task, and then the software would automatically create an efficient questionnaire design.

We thought there should be a significant advantage in letting each respondent have a unique experimental design, as we had in the aircraft study. At that time we hadn't heard of Hierarchical Bayes, and had not yet considered providing software for Latent Class analysis. At least initially, analysis would be done by aggregating data from many respondents. Although most previous conjoint applications had focused on main effects, we saw the possibility that this software might be able to handle interactions as well. The trick would be to let each respondent have a unique experimental design so that, in aggregate, the data would be informative about any contrast including all interactions between attributes taken two at a time.

Our initial method of questionnaire design was called "complete enumeration." It strove to create efficient designs individually for each respondent using a sequential method. The first concept in the first choice task was chosen at random, but subsequent concepts were chosen sequentially to ensure that, for each pair of attributes, every pair of levels appeared with as nearly the same frequency as possible. This also led to well-balanced designs because each level of an attribute appeared with approximately the same frequency, and no level was repeated more than necessary in any choice task.

In those early days computers were not nearly as fast as today. If there were many attributes and levels, the computational burden of evaluating every possible alternative at each stage would be considerable, and the respondent's wait for the next question could be intolerable. Accordingly, we also provided a "shortcut" method. Unlike complete enumeration, which kept track of co-occurrences of all pairs of attribute levels, the shortcut method considered attributes one-at-a-time. Shortcut designs were not as efficient as those using the complete enumeration method, but when pooling data among many respondents for aggregate analysis they were nearly as good. Joel Huber has reported that he and Klaus Zwerina tested randomized designs and found that for large designs, say, greater than 50 choice sets, they were 95% as efficient as optimal designs.

At the time, some other researchers' designs were "alternative-specific" and supported conditionality among attributes. For example, if one attribute was "car" vs. "bus", another attribute further describing "car" might be miles per gallon, and another attribute describing "bus" might be average waiting time until next bus. Or, as another example, in a pricing study the prices for single bottles of beverage might be less than a dollar, while prices for six-packs might be greater than a dollar. To provide this kind of flexibility in CBC designs, we introduced "prohibitions", which allowed the user to specify pairs of attribute levels that would never appear together, such as the average waiting time for the next car, or a single bottle of soft drink for several dollars. These prohibitions were detrimental to design efficiency, but were necessary to keep the questions meaningful. It was not until later years that CBC fully supported alternative-specific designs.

Initial Methods of Data Analysis

Although CBC can produce good data, it is not an efficient way to obtain data. The respondent should process the specifications of several concepts before providing each response, and that response only indicates his or her first choice, with no information about preferences among the remaining concepts. It was seldom possible to ask enough questions of a single individual to provide accurate estimates of his or her partworths. Therefore, in the first few years, CBC data analysis was usually done in aggregate, by pooling answers from many respondents.

The fact that CBC designs were well-balanced and unique for each individual conferred a valuable benefit. Within an attribute, every level was presented nearly the same number of times, and for each pair of attributes, every pair of levels was presented nearly the same number of times. Therefore, a simple counting procedure could produce estimates of average partworths for all main effects and interactions. We could count the proportion of those times when each level appeared that it had been in the chosen concept. That proportion measured the relative strength of preference for that level, and its logarithm was an estimate of the partworth that would be estimated by logit analysis. Similar computations could be made for interactions, using combinations of levels. These kinds of analysis were done automatically for main effects and two-way interactions.

Our choice of the word “randomized” to describe CBC’s designs may have been unfortunate. Some researchers trained in fixed designs recoiled at the idea, misunderstanding what we were doing, assuming that all respondents received the same randomized design rather than different ones. One reviewer took us to task for using what he regarded as a haphazard process in our designs. Perhaps it would have been better if we had used the word “stochastic” instead of “randomized.” But it is clear that choosing randomized designs was a good thing for CBC. The resulting designs are nearly optimal for D-efficiency when all attributes have the same number of levels, and can often beat completely orthogonal designs when attributes have different numbers of levels. Primarily, however, we have helped CBC users avoid the possibility of inadvertently choosing a fixed design which turned out to be defective in some way.

In its first few years CBC was used in many pricing studies employing just three attributes: brand, package size, and price. The automatic “counting” analysis of two-way interactions frequently showed brand x price interactions to be significant. Later, when Hierarchical Bayes became available and permitted individual-level analysis, interactions were typically found to be less important, often being due to aggregating unlike individuals.

There are some situations for which a counting analysis is adequate by itself. Other situations require simulators to answer what-if questions about the data. Such simulators require partworths, normally developed using multinomial logit analysis. Accordingly, the first version of CBC included the capability of doing aggregate logit analysis, as well as a simulator for answering what-if questions.

Improved Methods of Data Analysis

It was soon clear that aggregate analysis did not provide completely satisfactory answers. Pooling respondents to estimate a single set of average partworths neglected heterogeneity in respondent preferences. Indeed, widespread interest in market segmentation is based on the presumption that we need to know about the distribution of preferences.

Fortunately, Latent Class analysis had recently been introduced to the field. The conceptual basis of Latent Class had been introduced in sociology by Paul Lasarsfeld in about 1950. However, a recent article by DeSarbo and Ramaswamy[1] proposed a Latent Class approach appropriate for our kind of data which seemed able to bring respondent heterogeneity back

[1] DeSarbo, W. S. and Ramaswamy, V. (1995), "Marketing Segmentation with Choice Based Conjoint Analysis," Marketing Letters, 6, 137-148.

into the picture. With generous help from Wayne DeSarbo, I programmed a version of their approach. Latent Class still remains an appropriate solution when the goal is to present a relatively small number of segments to management. However, its assumption of homogeneity within segments meant that it did not do well predicting individual choices or choice shares.

Although partworths for segments provided a considerable improvement over just overall averages, that seemed only a half-way solution to the problem of retaining heterogeneity. It would have been much more satisfying to provide estimates of individual partworths. Latent Class analysis was able to something like that. Since each individual had an estimated probability of belonging to each class, a set of partworths could be estimated for each individual by probability-weighting the partworths for each class. Since the weights were probabilities, they were all positive and summed to unity. That meant that the estimates for each individual necessarily lay ***between*** estimates for the various segments.
Imagine three points in a plane, representing three Latent Class's partworths, with a triangle connecting those three points. Individual partworth estimates could be represented by points lying ***within*** that triangle. This way of estimating individual partworths seemed unnecessarily restricted. There seemed no good reason why the points for individuals should all lie between the points for segments, and we sought another way to use the Latent Class solution that would permit estimation of individual points not constrained to lie within the area bounded by the segment points

One approach to overcoming this problem was what we called ICE, for "Individual Choice Estimation". ICE also conceptualized each individual's partworths as a weighted combination of the Latent Class partworths, but with ICE the weights were unrestricted in range (i.e. could be either negative or positive). This meant that individuals' points could lie ***outside*** of the space between the segment points. For three segments, for example, an individual's point could lie anywhere on the plane determined by those three points. The weights for each individual were determined by logit analysis of his or her own choice data. Although responses to a few choice sets were not enough to estimate partworths all by themselves, they did seem adequate to estimate a small number of weights to be applied to Latent Class partworths.

We did several studies comparing ICE to results from aggregate logit and to partworths obtained by probability-weighting the Latent Class segment partworths. ICE appeared to do a better job than either of those. If it were not for the emergence of Hierarchical Bayes analysis, ICE might be in use today. However, just as we were concluding that ICE provided the best available way to capture heterogeneity in CBC data, along came Hierarchical Bayes (HB). And HB was better than ICE.

The HB Revolution

I owe a big debt to Greg Allenby for introducing me to Hierarchical Bayes analysis. He taught a workshop on HB at an ART Forum in the late 1990s, and distributed notes including a detailed description of an HB routine. I used that information to produce a program in the C language and tried it on some CBC data. The partworths estimated by HB appeared to be better in every respect than those estimated by other methods.

I have commented above that CBC is an inefficient procedure for collecting data, because the respondent should consider a lot of information before making each choice. As a result, our early attempts to estimate partworths necessarily operated at the aggregate rather than the individual level. Latent Class estimation relaxed this requirement somewhat, and ICE permitted us to get down to the individual level, but the quality of those estimates was not quite as good as estimates from HB. HB has overcome a basic limitation of CBC, and it is now possible to get reasonable individual estimates even when each respondent answers only a handful of choice questions. This is good for researcher and respondent alike. Indeed, HB expanded the usefulness of CBC so greatly that CBC is now the most frequently used of all conjoint methods. Based on an annual user feedback survey, we estimate that CBC surpassed the usage frequency of ACA in about the year 2000.

D-Efficiency and the Logit Rule

In 1996 Huber and Zwerina published a paper[2] which clarified some important issues concerning efficiency of choice designs. For designs estimated by linear regression, D-efficiency depends only on the characteristics of the design itself, and not on response data. However, for logit analysis of choice data, D-efficiency depends on both the design and the partworths to be estimated. Huber and Zwerina showed that efficiency of choice designs could be improved if the concepts within each choice set were selected so as to be more nearly equal in attractiveness.

This led to a series of experiments at Sawtooth Software, starting in 2003. Our basic idea was to use information from each respondent's earlier choice tasks to estimate his or her partworths, and then choose the next choice task so its alternatives were more nearly equal in attractiveness. Three studies were done, all using algorithms which sought to maximize D-efficiency, each in a more effective way than before.

In all three studies we were successful at the experimental manipulation; that is to say, we were able to produce designs with D-efficiency considerably greater than found with standard CBC. In each study we included several holdout choice sets, to see whether we could improve hit rates in predicting holdout choices. The first study seemed to suggest that greater D-efficiency did indeed lead to slightly better predictions, but subsequent studies failed to confirm the benefit of increasing D-efficiency.

These results presented a puzzle. ***If*** respondents are using logit models to make choices, then more efficient designs ***must*** do a better job of estimating their partworths. It would follow that

[2] Huber, Joel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," Journal of Marketing Research 33, (August), 307-317.

we should be more successful at predicting holdout choices with more efficient designs, but we were ***not*** more successful. We seemed to have found evidence that ***most respondents do not use logit models in making choices in CBC.***

About this same time, two papers (Gilbride & Allenby as well as Hauser, Dahan, Yee, & Orlin)[3] showed that simpler models could do a good job of explaining respondent choices. Following the second of these papers, we reexamined several recent data sets. For a typical data set we found that 80 percent of respondents had answers that could be accounted for by the presence or absence of only three or fewer attribute levels. It seemed likely that most respondents were ***not*** using logit rules to respond to choice tasks. Instead, each respondent might be paying attention only to a small number of "must have" attribute levels.

There is further evidence that respondents use strategies that are simpler than logit. When response times are measured, the average time required to study the alternatives in a choice task and make a choice is typically just ten to fifteen seconds. It seems unlikely that respondents could truly consider all the attribute levels for several alternatives in so short a time.

The Lure of Minimal Overlap

Huber and Zwerina described four characteristics of choice designs that lead to D-efficiency:

> **Level Balance:** The levels of an attribute occur equally often.
>
> **Orthogonality:** For any two attributes, all pairs of levels occur equally often.
>
> **Minimal Overlap:** Attribute levels are repeated within a choice set as few times as possible.
>
> **Utility Balance:** Within a choice set, alternatives are equally attractive.

These characteristics are contradictory to some extent, because utility balance cannot usually be achieved without trade-offs in the others. CBC designs created by the complete enumeration method do a good job with the first three characteristics, but ignore utility balance. Both of the design methods offered initially in CBC, complete enumeration and the shortcut method, attempt to minimize overlap.

Given this new information suggesting that respondents might not be using logit models in their choices, we realized that minimal overlap might not be a good characteristic of choice designs.

[3] Gilbride, Timothy and Greg M. Allenby (2004), "A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," Marketing Science, 23, 3, (Summer) 391-406.
Hauser, John R., Ely Dahan, Michael Yee, and James Orlin (2006) "'Must Have' Aspects vs. Tradeoff Aspects in Models of Customer Decisions," Sawtooth Software Conference Proceedings, 169-181.

Suppose three brands are being studied, and there are three alternatives in each choice set. Minimal overlap will be achieved if all three brands appear in each choice set. This will lead to higher D-efficiency than would be achieved with more overlap, but consider how this simplifies the respondent's job. If there is a single "must have" brand, then all the respondent has to do is choose that alternative, without even considering the other attributes. Minimal overlap may help respondents get through choice questionnaires with less thought, not a desirable feature.

We had become aware earlier that minimal overlap, while good for efficient measurement of main effects, was destructive to the measurement of interactions. Since we wanted to strengthen CBC's ability to handle interactions, we had decided to add two new design methods:

> **Random Method:** The random method employs random sampling with replacement for choosing concepts, which permits an attribute to have identical levels across all concepts, but it does not permit two identical concepts within the same task.
>
> **Balanced Overlap Method:** This method is a middling position between the random and the complete enumeration strategies. It permits roughly half as much overlap as the random method. It keeps track of the co-occurrences of all pairs of attribute levels, but permits more level overlap within the same task.

These two methods of constructing choice sets are less efficient for measuring main effects than the original two methods, but more efficient at measuring interactions. More to the point, however, they make it harder for respondents to adopt the simple strategy of always choosing the single concept with a particular attribute level. Given evidence that respondents tend to simplify their tasks when possible, the Balanced Overlap method is likely to provide more informative data.

Adaptive CBC

I have already mentioned that, following the Huber and Zwerina article, we conducted a series of studies aimed at developing more D-efficient versions of CBC. We were able to produce greater D-efficiency, but the resulting partworth estimates were no better at predicting holdout choices. In addition, more recent evidence had become available that respondents may use simplified ways of responding, rather than adhering to a logit rule.

We concluded that we had probably been on the wrong track in trying to design questionnaires with greater statistical efficiency under the logit model. What was needed, instead, was a way of permitting the respondent to express the need for any "must have" features and reject any "absolutely unacceptable" features, and then to focus on trade-offs among the remaining features illustrated by concepts tailored to his or her preferences. ACBC ("Adaptive CBC"), which was introduced in 2009, does exactly that. Additionally, ACBC focuses on product concepts relevant for the respondent by providing a "Build Your Own" section, and then exploring alternatives similar to the product described by the respondent.

We have compared ACBC with standard CBC in several studies, and ACBC seems to provide richer data, and is better able to predict respondent choices in holdout tasks. ACBC takes longer than standard CBC, but it provides more information. And respondents appear not to mind the longer interview time, rating ACBC more favorably than standard CBC in terms of letting them express their preferences effectively and being less boring.

ACBC has been gaining popularity since its release two years ago. Our user feedback surveys indicate that the ratio of ACBC to CBC projects was about 1:11 in the first year, and about 1:7 at the end of the second year. Use of ACBC will doubtless continue to increase, particularly after we provide the capability of doing alternative-specific designs.

What's Next?

Since its beginning about 20 years ago, CBC has grown to become the most widely used conjoint technique. One of its strengths is that choices are easy for respondents to make, so almost everyone is capable of completing a CBC interview. Another strength is that the act of choosing an alternative is similar to what buyers do when considering purchases, so respondents are likely to find CBC interviews interesting, and researchers are more likely to regard the resulting data as meaningful.

It seems likely that CBC will continue to be an important tool. In the first version of CBC, we constrained the number of attributes to a small number, because we doubted that respondents would be able to handle choice tasks effectively if there were many attributes.   CBC may continue as the most effective way to handle projects with few attributes, such as brand-package-price studies. Or perhaps CBC will continue to develop in some other direction.

CBC has had a great past. It will be interesting to see what it does in the future!