Sawtooth Software

RESEARCH PAPER SERIES

Getting the Most from CBC

Rich Johnson and Bryan Orme,
Sawtooth Software, Inc.

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

Getting the Most from CBC

Rich Johnson and Bryan Orme, Sawtooth Software

We originally published an earlier version of this article spanning two issues of our newsletter, Sawtooth Solutions, in Fall 1996 and Spring 1997. We've condensed the two articles and updated some of the content to reflect recent developments.

When designing CBC, we tried to make many of its features automatic, so users could accomplish their goals by accepting built-in defaults. However, CBC users have had questions about several issues, and these are our recommendations about them.

Using prohibitions: CBC lets you specify that certain combinations of levels should not appear in the questionnaire, such as a luxury product at an economy price. Prohibitions should be used only rarely, to avoid showing hypothetical products that would be seen as completely absurd. There are several reasons to avoid using prohibitions:

> (1) CBC lets you specify unique price ranges (Conditional Pricing) for different combinations of levels, such as brand and package size, so prohibitions are not required to ensure reasonable prices.
>
> (2) For some situations in which extreme prohibitions between attributes seem to be necessary to reflect reality, CBC's Advanced Design Module can offer a more sound approach with its "Alternative-Specific" design strategy.
>
> (3) Respondents are usually able to deal with product concepts more "extreme" than currently available in the market, and permitting such combinations not only increases the efficiency of estimation of their utilities, but also let you estimate "what might be" as well as "what is."
>
> (4) Too many prohibitions, or a particular pattern of them, may make it impossible to create a good design, and may undermine your ability to analyze the data. Although all prohibitions decrease statistical efficiency, CBC provides a way to test a design before you go to the field to ensure that prohibitions will not have catastrophic effects. If you do use prohibitions, **test your design**! If the test design report shows asterisks for the standard errors of any attribute level, or if it reports "design deficient" you must re-consider the prohibitions. As a further test, you may want to generate dummy response data for alternative designs, run logit, and examine the resulting standard errors.

Numbers of attributes and levels: CBC permits a maximum of 10 attributes (30 with the Advanced Design Module), and a maximum of 15 levels per attribute (or up to 100 levels per attribute with CBC/Web's Advanced Design Module). But, as with other conjoint methods, you get the best results if you keep things simple. Don't include unnecessary attributes or levels just because there's room for them. Many CBC studies

use only three or four attributes. You may have to use many levels for "categorical" attributes like Brand or Package Type, but there's seldom any reason to have more than 5 levels for quantitative attributes like Price. It's usually better to have more data at each price point than to have thinner measurements at more price points, particularly if you're interested in interactions. The interaction between two 9-level attributes involves 64 logit parameters, but the interaction between two 5-level attributes requires only 16.

Determining sample size: In a paper presented at the 1996 ART forum ("How Many Questions Should You Ask in Choice-Based Conjoint Studies?" available within our Technical Papers library at www.sawtoothsoftware.com) we showed that the statistical gain from increasing the number of choice tasks per respondent was similar to the gain from a proportional increase in the number of respondents. This conclusion was based on the assumption of aggregate (pooled) logit analysis, which is now not used as often as Latent Class or hierarchical Bayes (HB) methods. Despite the shortcomings of aggregate logit analysis, assuming a pooled model permits the use of simple statistical tests to determine how sample size relates to precision of aggregate part worth estimates. Although this approach is now somewhat dated, we include it as "food for thought" when considering sample sizes and CBC.

1. Count the number of "cells" in the largest interaction you want to measure. For example, with the interaction of two 9-level attributes, there are 81 cells. One way to think about CBC analyses is that you want to estimate the proportion of times that concepts containing levels identifying each cell are chosen.

2. Determine how many concepts will be shown altogether, which is the number of respondents times the number of choice tasks per respondent times the number of concepts per task.

3. The average number of occurrences of each cell will be equal to the total number of concepts shown, divided by the number of cells. Call this quotient **n**.

4. Ignoring choices of "None," the average probability of a concept being chosen is 1 over the number of concepts per task. Call this **p**.

5. The standard error of a proportion is $\sqrt{p(1-p)/n}$

(Here, as in many other statistics used in aggregate logit analysis, we assume that the choices made by each respondent are independent of one another.)

For example, consider a "typical" CBC study, with 300 respondents, each with 10 choice tasks, and with 5 concepts in each task. The total number of concepts shown will be 15,000. If the largest interaction we want to measure is 5 x 5, the average number of occurrences of each cell will be 15,000 / 25 = 600. The average probability of a concept's being chosen will be 1/5 = .2, so the average standard error will be sqrt [.2 * .8 / 600] = .016, and the 95% confidence interval will be about +/- 1.96 * .016, or +/- .03.

If we were interested in a 9x9 interaction, the number of cells would be 81 rather than 25, and the number of respondents required for equivalent precision would be about three times as large. If we were only interested in main effects, the maximum number of cells might be 5 rather than 25, and a sample size only a fifth as large could yield equivalent precision for those estimates.

The "typical" CBC study is like our example, in which the total number of concepts divided by the number of cells of interest is about 600. If separate estimates are desired for several segments, then adequate samples should be included for each of them. Finally, remember that you can get about the same increase in precision from proportional increases in the number respondents, or the number of tasks per respondent, for questionnaires with up to 20 tasks.

Sample size and HB

The previous discussion on sample size assumed aggregate (pooled) analysis, which is not used as often today as Latent Class and HB methods. Latent Class and HB can model respondent heterogeneity (differences in preferences across people), resulting in more accurate market simulations. Using Latent Class or HB significantly reduces IIA problems (discussed later in this document).

HB analysis yields individual-level estimates of part worths, and so the guidelines presented in the previous section are no longer so directly applicable. The traditional considerations for ratings-based conjoint or ACA must be considered: enough tasks should be asked of each respondent to permit stable individual-level estimates, and enough respondents should be included to reduce sampling error. For more guidance regarding sample sizes in conjoint analysis studies, please refer to an article entitled "Sample Size Issues for Conjoint Analysis Studies" available within the Technical Papers library at www.sawtoothsoftware.com.

Reporting results: counts vs. simulations: If you don't use prohibitions, CBC produces designs in which each attribute varies independently of the others. This means that you can measure the effect of an attribute simply by observing the proportion of times concepts are chosen when they have each level. Similarly, two-way interactions can be evaluated by seeing how often concepts with each pair of levels are chosen. We call this the "counting" approach. You can also do logit, Latent Class or HB analysis to estimate part worth utilities, which you can then use in market simulations. If you have not used prohibitions, counts will produce results similar to average part worths from these utility estimation methods.

For simple questions, such as obtaining price sensitivity curves for specific brands, the counting approach is sometimes adequate. Since no complicated analysis is required, results are easy to communicate to others who are not market researchers. However, other objectives may require part worth estimation and market simulations. For example, if you want to simulate a particular product's showing in a specific competitive market,

including its share as its price changes or competitors' prices change, the simulation approach is more appropriate.

Including "None": CBC provides the option of letting the respondent choose "None," or another constant alternative such as "I'd continue buying my usual product." One issue is whether to include "None" in the questionnaire, and a separate issue is whether to include "None" in the analysis.

We think it is usually a good idea to include the "None" option in the questionnaire, for these reasons:

- It makes the choice tasks more realistic, because that option is usually available when shopping.

- It makes the experience more pleasant for the respondent, who is not forced to select an unacceptable alternative.

- It improves the quality of the data, by letting respondents screen themselves out of questions containing only alternatives they would never consider.

However, we recommend you conduct a pretest to ensure that None is not chosen too often. When respondents choose None, very little information is gained for refining the estimates of the attribute levels. If the None incidence is too high (a typical range is 5%-15%), perhaps you are interviewing the wrong people or you may need to review the attribute levels.

Some researchers like to include a "None" category in simulations, as an aid in estimating how category volume would expand or shrink as products become more attractive. We recommend caution if doing this, for these reasons:

- CBC's estimate of how many respondents should choose "None" depends on the number of alternatives in the choice tasks. If you use aggregate logit analysis and conduct a simulation with a different number of products, your estimates will likely not be correct (due to IIA assumptions). However, this problem is significantly reduced if using Latent Class or HB. In a recent data set analyzed using HB, we found that the resulting "None" share was appropriate when the number of products in a market simulation was varied from 2 to 4 alternatives (as compared to holdout choice tasks which also varied the number of product alternatives from 2 to 4).

- Although choices of "None" are probably indicative of disliked alternatives, there is little reason to believe their frequency will accurately reflect the actual proportion of respondents refusing to purchase products in the real world.

In summary, we usually suggest including the "None" option in choice tasks, but then paying less attention to or even neglecting it in the analysis.

Calibrating CBC results to market shares: CBC results usually differ from actual market shares. This is not surprising, since market shares are influenced by product distribution, brand awareness, out-of-stock conditions, point-of-sale promotions, imperfect buyer knowledge, and many other factors not captured in conjoint measurement.

Researchers are often motivated to adjust or "calibrate" simulation results to look like market shares. We suggest not doing so, because no matter how carefully choice results are calibrated to the market, the researcher will one day be embarrassed by differences that remain. However, if the pressure to do so is too great to resist, there are two ways CBC results can be adjusted to more closely mimic market shares.

CBC utilities are scaled automatically to reflect the amount of random error in respondents' choices. You can over-ride that scaling by specifying a scaling parameter (exponent). Larger values than the default of 1.0 will create greater variation among shares of choice, making large simulated shares even larger, and small shares even smaller. Smaller values of the parameter will create less variation among simulated shares of choice, in the limit making them all equal.

Market shares are often "flatter" than choice shares, because they are affected by additional sources of random noise. If that is the case, you may be able to approximate market shares more closely with a scaling parameter of less than 1.0. Beware, however, that such an adjustment will also make your results less sensitive to changes, including pricing changes.

Sometimes, market shares reflect ***too much*** variation, for example when the largest product has nearly 100% geographic distribution but smaller products do not. In that case we suggest **not** adjusting the scaling parameter, which could make your results too sensitive to pricing changes.

In general, if you must make simulated shares look like market shares, we suggest using "external effects" which will adjust share levels to be like market shares. However, introducing external effects will affect a product's sensitivity to change (e.g. price elasticity). External effects may be calculated for each simulated product by dividing the target share by the simulated share of choice.

IIA and the Red Bus/Blue Bus problem: The Share of Preference model in the market simulator, suffers from "IIA," which is shorthand for "Independence from Irrelevant Alternatives." The basic idea of IIA is that the ratio of any two products' shares should be independent of all other products. This sounds like a good thing, and at first, IIA was regarded as a beneficial property.

However, another way to say the same thing is that an improved product gains share from all other products in proportion to their shares; and when a product loses share, it loses to others in proportion to their shares. Stated that way, it is easy to see that IIA implies an

unrealistically simple model. In the real world, products compete unequally with one another, and when an existing product is improved, it usually gains most from a subset of products with which it competes most directly.

Imagine a transportation market with two products, cars and red busses, each having a market share of 50%. Suppose we add a second bus, colored blue. An IIA simulator working from aggregate logit part worths would predict that the blue bus would take share equally from the car and red bus, so that the total bus share would become 67%. But it's clearly more reasonable to expect that the blue bus would take share mostly from the red bus, and that total bus share would remain close to 50%. Indeed, the IIA problem is sometimes referred to as the "red bus, blue bus problem."

Our market simulator offers a "first choice" model, which may be used if you have individual-level part worths and avoids the IIA assumption entirely. If you do a first choice simulation and add a product identical to an existing product, those two products will get the same total share of first choices as either would alone. However, the first choice model is usually not satisfactory for another reason: it tends to overstate results for the most popular products.

There are a few ways to reduce the problems due to IIA. The most important way to reduce IIA problems is to use Latent Class (with a high dimensionality solution) or, preferably, HB. A second way is to use a market simulation model that does not assume IIA. In many situations, first choice simulations are "too extreme" to reflect proper scaling of market shares relative to the market. However, if you discover that the first choice simulation produces appropriately scaled shares, then this model would be appropriate. The more recent Randomized First Choice simulation method controls IIA issues, and is generally recommended because it can be "tuned" to reflect flatter or steeper share estimates, as well as greater or lesser control for IIA. Our market simulator also offers an older "Correction for Product Similarity" model that penalizes products in proportion to their similarity to others. That correction is generally eclipsed by the more accurate Randomized First Choice model. It is available within our simulator for historical purposes only.

Many of our users do "sensitivity analyses" by varying a product up and down on each attribute in turn. When testing sensitivity in this way, one can sometimes run into difficulties if using methods that correct for product similarity, even Randomized First Choice. It can give misleading answers if all products are initially simulated as having the same mid-level price. If one product is then moved to the next adjacent price point, it receives extra consideration by being unique relative to the others. This adjustment for similarity can sometimes cause a biased "kink" in the demand curve between the "average" price point and the adjacent levels (this "kink" is even more pronounced if using the older Correction for Product Similarity method). If you notice an uneven kink in the demand curve due to the issues noted above, we recommend using HB part worths and the Share of Preference model. One might still worry about the IIA properties of this model, but we've generally have found the Share of Preference model to perform quite well as long as individual-level part worths are used. For an explanation of how

capturing heterogeneity reduces IIA problems, please see an article entitled "The Benefits of Accounting for Respondent Heterogeneity in Choice Modeling" in our Technical Papers library at www.sawtoothsoftware.com.

Another common situation involves simulating the effect of line extensions. If including two (or more) only slightly different versions of a product of interest, but only a single version of other products, it is critical to control IIA. If not, it gives an artificial edge to the product with two versions.

Finally, models subject to IIA difficulties are not very good at measuring cross-elasticities. They can do quite a good job of measuring the effect of a price change on that product's ***own*** share, but they are not very good at measuring the effect of a price change on ***other*** products' shares. The reason, again, is that IIA restrictions require that a product's interactions with others will be proportional to their shares. If using a high-dimensionality Latent Class solution or individual-level part worths, this problem is significantly reduced.