Sawtooth Software

RESEARCH PAPER SERIES

Comparing Hierarchical Bayes Draws and Randomized First Choice for Conjoint Simulations

Bryan K. Orme and Gary C. Baker,
Sawtooth Software, Inc.,
2000

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

Comparing Hierarchical Bayes Draws and Randomized First Choice for Conjoint Simulations

Bryan Orme and Gary Baker,
Sawtooth Software, Inc.

Introduction

Conducting market simulations is one of the most valuable uses of conjoint analysis data. Market simulations transform raw conjoint part-worths (which to a non-researcher can seem quite esoteric) to the more managerially satisfying model of predicting buyer choices for specific market scenarios.

The last few years have seen important new developments for estimating conjoint part-worths and simulating shares of preference. Foremost among the advances in part-worth estimation in our opinion is the use of hierarchical Bayes (HB) to estimate individual-level part-worths from conjoint or choice data. Another recent advancement has been the introduction of Randomized First Choice for conducting market simulations (Orme 1998, Huber *et al.* 1999) and its general availability within Sawtooth Software's conjoint analysis systems.

The typical application of conjoint analysis has involved estimating many independent parameters (attribute level part worths) from only marginally more observations (questions/tasks). Despite this, the results (especially for predicting aggregate shares of preference) typically have been quite useful.

And then HB became available. HB is a very effective "data borrowing" technique that stabilizes part-worth estimates for each individual using information from not only that respondent, but others within the same data set. HB generates multiple estimates of the part-worths for each respondent, called *draws*. These multiple draws can be averaged to create a single vector of part-worths for each respondent (*point estimates* of the part-worths). One can also use those draws to estimate the variances and covariances of part-worths within each respondent. Another potential application is to use the draws themselves, rather than point estimates, in market simulations. Reviewing the theory and mathematics behind HB are beyond the scope of this paper. For the interested reader, we suggest the CBC/HB Technical Paper (Sawtooth Software, 1999).

Over the years, two main simulation models have been applied to part-worth data: First Choice and Share of Preference (logit). The First Choice model, while immune to IIA, is typically too extreme (and not tunable for scale). The Share of Preference model, while tunable, suffers from IIA. Randomized First Choice (RFC) adds back random variation to the point estimates of the part-worths during simulations. RFC is appropriate for aggregate (logit) or disaggregate part-worths (e.g. ACA, traditional full-profile conjoint (CVA), Latent Class or even self-explicated utilities). Each respondent's (or group's) point estimates are sampled multiple times during simulations, with different random variance added each time. The utility of alternatives is computed at each iteration (draw) and choices are assigned applying the First Choice rule. On

the surface, this approach resembles using HB draws in simulations. Both techniques reflect uncertainty (error distributions) about the part-worths and simulate multiple choices per respondent (or group).

In this paper, we compare the use of HB draws and RFC. Our findings show that using HB draws in simulations seems to work well, but applying RFC to individual-level point estimates of part-worths works even better. We also discuss two potential biases when using HB draws in simulations: a *reverse number of levels effect*, and an *excluded level effect*. These biases may in part explain why simulating using draws was not as successful as applying RFC to point estimates for our data set.

Random Effects Theory

Before we continue, we should review the basic random effects model. The random effects model (McFadden 1973) takes the following form:

$$U_i = X_i (\beta) + \varepsilon_i$$

where U_i = utility of alternative i
X_i = row vector of independent variables (attribute level codes) associated with alternative i
β = vector of part-worths
ε_i = an error term

In fact, if ε_i is distributed as *Gumbel and the First Choice rule is applied in simulations, the expectation (given a very large number of draws) is identical to the logit simulation model. Adding larger error variance to the utility of each alternative is equivalent to applying a smaller "scale factor" and share predictions are "flattened." Adding less error makes the share predictions "steeper." Adding zero error to the utility of each alternative is equivalent to the First Choice model. During the remainder of this paper, the reader should keep in mind this inverse relationship between error and the resulting scale of the predicted shares.

The next three sections are an introduction to Randomized First Choice and are taken (with a few minor modifications and additions) from the CBC v2.0 manual (Sawtooth Software, 1999). Those familiar with RFC may choose to skip these sections.

Previous Simulation Methods

The First Choice model (maximum utility rule) has a long history in conjoint analysis for simulating shares of preference among competitive product concepts. It is intuitively easy to understand and is immune from IIA (Red Bus/Blue Bus) difficulties. However, it also often does not work very well in practice. The share estimates usually tend to be too "steep" relative to

* The Gumbel (extreme value) distribution is a double negative exponential distribution, drawn by taking ($Y = -\ln(-\ln x)$), where x is a rectangularly distributed random variable ($0<x<1$).

shares in the real world. The standard errors of the simulated shares are much greater than with logit (Share of Preference) simulations, since product preference is applied as an "all or nothing" 0/1 rule. Moreover, the notion that a respondent who is "on the fence" with respect to two alternatives will make a sure choice is simplistic and counter-factual (Huber *et al.* 1999).

> ***Main Point #1: First Choice share predictions are usually too extreme and are not tunable. But, they avoid IIA problems.***

The Share of Preference (logit) simulation model offers a way to tune the resulting shares to the desired scaling. It also captures relative information about the value of *all* product alternatives rather than just the best one, thereby increasing the precision of the simulated shares. However, the model is subject to IIA (Red-Bus/Blue-Bus) problems. Within the unit of analysis, cross-elasticities and substitution rates among products are assumed to be constant. This drawback can be quite damaging—especially for aggregate models (i.e. aggregate logit or Latent Class).

> ***Main Point #2: Share of Preference share predictions can be tuned and have greater precision than First Choice shares. But, they have IIA problems.***

Randomized First Choice

The Randomized First Choice (RFC) method combines many of the desirable elements of the First Choice and Share of Preference models. As the name implies, the method is based on the First Choice rule, and helps significantly resolve IIA difficulties. As with the Share of Preference model, the overall scaling (flatness or steepness) of the shares can be tuned.

Most of the theory and mathematics behind the RFC model are nothing new. However, to the best of our knowledge, those principles had never been synthesized into a generalized conjoint/choice market simulation model. RFC, suggested by Orme (Orme 1998) and later refined by Huber, Orme and Miller (Huber *et al.* 1999), was shown to outperform all other Sawtooth Software simulation models in predicting holdout choice shares for a data set they examined. The holdout choice sets for that study were designed specifically to include identical or near-identical alternatives.

Rather than use the part-worths as point estimates of preference, RFC recognizes that there is some degree of error around these points. The RFC model adds unique random error (variation) to the part-worths and computes shares of preference using the First Choice rule. Each respondent is sampled many times to stabilize the share estimates. The RFC model results in an automatic correction for product similarity due to correlated sums of errors among product alternatives defined on many of the same attributes. To illustrate RFC and how correlated errors added to product utilities can adjust for product similarity, consider the following example:

Assume two products: A and B. Further assume that A and B are unique. Consider the following product utilities for a given respondent:

	Avg. Product Utilities
A	10
B	30

If we conduct a First Choice simulation, product B captures 100% of the share:

	Avg. Product Utilities	Share of Choice
A	10	0%
B	30	100%

However, let's assume that random forces come to bear on the decision for this respondent. Perhaps he is in a hurry one day and doesn't take the time to make the decision that optimizes his utility. Or, perhaps product B is temporarily out-of-stock. Many random factors in the real world can keep our respondent from always choosing B.

We can simulate those random forces by adding random values to A and B. If we choose large enough random numbers so that it becomes possible for the utility of A sometimes to exceed the utility of B, and simulate this respondent's choice a great many times (choosing new random numbers for each choice replication), we might observe a distribution of choices as follows:

	Avg. Product Utilities	Share of Choice
A	10	25.00%
B	30	75.00%

(Note: the simulation results in this section are for illustration, to provide an intuitive example of RFC modeling. For this purpose, we assume shares of preference are proportional to product utilities.)

Next, assume that we add a new product to the mix (A'), identical in every way to A. We again add random variation to the product utilities so that it is possible for A and A' to be sometimes chosen over B, given repeated simulations of product choice for our given respondent. We might observe shares of preference for the three-product scenario as follows:

	Avg. Product Utilities	Share of Choice
A	10	20.0%
A'	10	20.0% (A + A' = 40.0%)
B	30	60.0%

Because unique (uncorrelated) random values are added to each product, A and A' have a much greater chance of being preferred to B than either one alone would have had. (When a low random error value is added to A, A' often compensates with a high random error value). As a

simple analogy, you are more likely to win the lottery with two tickets than with one.

Given what we know about consumer behavior, it doesn't make sense that A alone captures 25.0% of the market, but that adding an identical product to the competitive scenario should increase the net share for A and A' from 25.0% to 40.0% (the classic Red Bus/Blue Bus problem). It doesn't seem right that the identical products A and A' should compete as strongly with one another as with B.

If, rather than adding uncorrelated random error to A and A' within each choice replication, we add the same (correlated) error term to both A and A', but add a unique (uncorrelated) error term to B, the shares computed under the First Choice rule would be as follows:

	Avg. Product Utilities	Share of Choice
A	10	12.5%
A'	10	12.5% (A + A' = 25.0%)
B	30	75.0%

(We have randomly broken the ties between A and A' when accumulating shares of choice). Since the same random value is added to both A and A' in each repeated simulation of purchase choice, A and A' have less opportunity of being chosen over B as compared to the previous case when each received a unique error component (i.e. one lottery ticket vs. two). The final utility (utility estimate plus error) for A and A' is always identical within each repeated First Choice simulation, and the inclusion of an identical copy of A therefore has no impact on the simulation result. The correlated error terms added to the product utilities have resulted in a correction for product similarity.

Let's assume that each of the products in this example was described by five attributes. Consider two new products (C and C') that are not identical, but are very similar—defined in the same way on four out of five attributes. If we add random variation to the part-worths (at the attribute level), four-fifths of the accumulated error between C and C' is the same, and only one-fifth is unique. Those two products in an RFC simulation model would compete very strongly against one another relative to other less similar products included in the same simulation. When C received a particularly large positive error term added to its utility, chances are very good that C' would also have received a large positive error term (since four-fifths of the error is identical) and large overall utility.

RFC Model Defined

We can add random variation at both the attribute *and* product level to simulate any similarity correction between the IIA model and a model that splits shares for identical products:

$$U_i = X_i (\beta + E_a) + E_p$$

where:

U_i	=	Utility of alternative i for an individual or homogenous segment at a moment in time
X_i	=	Row of design matrix associated with product i
β	=	Vector of part-worths
E_a	=	Variability added to the part-worths (same for all products in the set)
E_p	=	Variability (i.i.d Gumbel) added to product i (unique for each product in the set)

Repeated draws are made to achieve stability in share estimates, computed under the First Choice rule. We used E_a error distributed as Gumbel, but a normal distribution could be used as well.

(Note that when the attribute variability is zero, the equation above is identical to the random effects model presented earlier, which is identical to the logit rule.)

In RFC, the more variation added to the part-worths, the flatter the simulations become. The less variation added to part-worths, the steeper the simulations become. Under every possible amount of attribute variability (and no product variability), net shares are split in half for identical products, resulting in no "inflation" of net share. However, there may be many market scenarios in which some share inflation is justified for similar products. A second unique variation term (E_p, distributed as Gumbel) added to each product utility sum can tune the amount of share inflation, and also has an impact on the flatness or steepness of the overall share results. It can be shown that adding only product variability (distributed as Gumbel) within the RFC model is identical to the familiar logit model (Share of Preference Model). Therefore, any degree of scaling or pattern of correction for product similarity ranging between the First Choice model and Share of Preference can be specified with an RFC model by tuning the *relative* contribution of the attribute and product variation.

The obvious question for the researcher is how much share inflation/correction for product similarity is justified for any given modeling situation. To answer this, holdout choice tasks that include some alternatives that are very similar alongside others that are quite unique should be included in the study and used for tuning the RFC model.

Using HB Draws in Simulations

As introduced earlier, hierarchical Bayes can estimate individual-level part-worth utilities for choice or conjoint experiments. HB is a computationally-intensive iterative process. After a period of "burn-in" iterations, convergence is assumed and the results of subsequent iterations are saved. One usually saves many iterations (draws) for each respondent. Point estimates of part-worths are computed for each individual as the average of the saved draws.

The analyst has the choice of using the point estimates or the multiple draws per respondent in market simulations. Indeed, using HB draws rather than point estimates has a great deal in common with RFC. The draws reflect error distributions around the average parameters for each individual. However, the variances and covariances of the parameters in HB are empirically estimated. In contrast, RFC assumes (within each respondent or unit of analysis) that the variances of the part-worths are equal and the covariances are zero.

> ***Main Point #3: Two simulation methods show promise for reducing the IIA problem while at the same time being tunable: using HB draws and RFC.***

Before comparing HB draws and RFC, we were faced with the question of how many draws to use for HB simulations and how many sampling replications to employ with RFC. The answer would affect the computational time required to perform the simulations for this paper. More importantly, we recognize that researchers face the same decision when analyzing real-world studies—and they are usually under greater time constraints than we were.

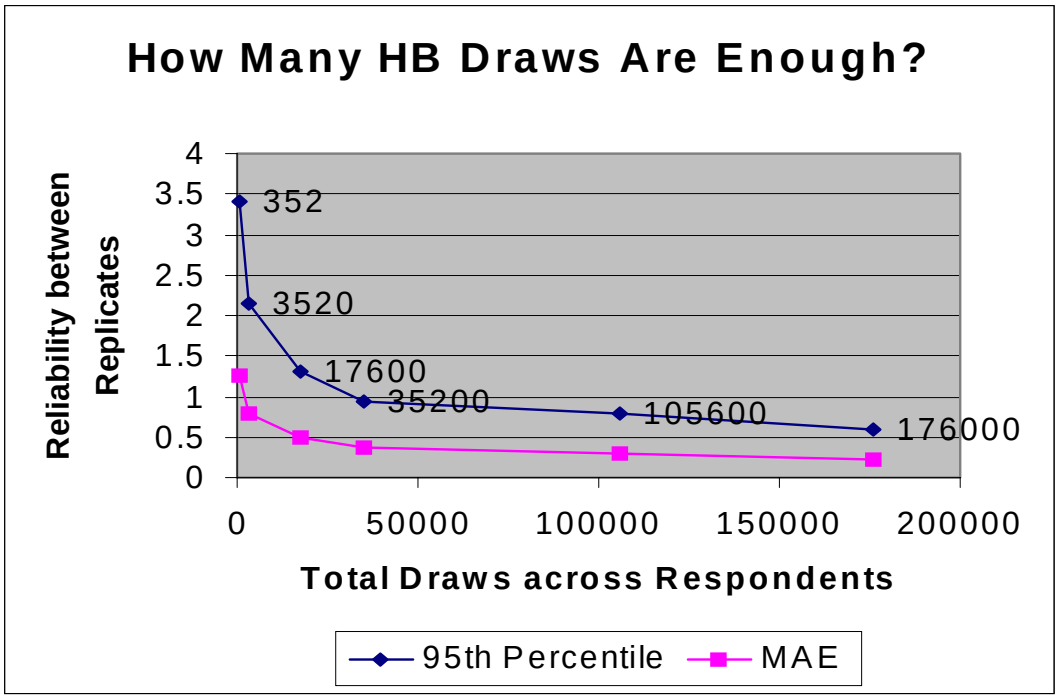
As background, the data set we used was collected by Huber, Orme and Miller in 1997. Shoppers were intercepted at shopping malls and asked to participate in a CBC study dealing with television sets. Three-hundred fifty-two respondents saw 18 randomly designed choice tasks followed by nine fixed holdout choice tasks. The design had six total attributes with 17 total levels. The holdouts were very carefully designed to have near utility balance. Also, each holdout choice task included five product concepts, two of which were either identical, or nearly identical. There were four versions of the fixed holdout choice tasks resulting in (4 versions)(9 tasks)(5 concepts per task) = 180 product concepts for which we could compare actual versus simulated shares of preference.

We ran HB for 35,000 burn-in iterations, then saved 1000 draws (skipping every tenth draw) per respondent, for a total of (352 respondents)(1000 draws) = 352,000 sets of 17 part-worths, resulting in a 57 Megabyte data file. To test the stability of simulated shares of preference given different numbers of draws per respondent, we simulated shares (First Choice rule) for the 180 holdout products using only the first draw for 352 respondents compared to the shares computed using only the 1000th draw. The difference was measured in terms of MAE (Mean Absolute Error). We repeated the analysis multiple times to stabilize the results (draw #2 versus draw #999, draw #3 versus draw #998, etc). We also sorted the MAE figures from worst (largest) to best (smallest) and recorded the 95th percentile MAE. We repeated that same analysis for 10 draws at a time per respondent, 50, 100, 300 and 500. The results are displayed in the table and graph below. For example, when using just one draw per respondent in simulations, the

simulated shares for holdout concepts differed on average by 1.25 share points between replicates, but 5% of the shares differed by 3.41 points or more.

Table 1

Draws per Respondent	Total Draws across All Respondents (n=352)	Mean Absolute Error between Replicates	95th Percentile Mean Absolute Error
1	352	1.25	3.41
10	3,520	0.80	2.16
50	17,600	0.49	1.31
100	35,200	0.38	0.95
300	105,600	0.30	0.78
500	176,000	0.22	0.59



With 35,200 draws, 95% of the shares between replicates differ by less than a share point. We think the data suggest that using more than about 50,000 total draws does not offer *practical* benefit in terms of stability of aggregate share estimates. Until computers become much faster, if the researcher is only interested in aggregate share predictions, we suggest using at least 30,000 but not much more than about 50,000 total draws (across all respondents). Much larger than that and the files become difficult to manage and processing time too much to swallow for "in-the-trenches" research. However, the constraint of file size assumes that the researcher wants to use HB draws during simulations, which we'll argue may not be the suggested approach.

Comparing HB Draws and RFC

We compared simulations from HB draws and RFC on point estimates by examining the predicted versus actual holdout shares. MAE (Mean Absolute Error) quantifies the difference between actual and predicted shares. For example, suppose we have three products with actual share percentages of 10, 30, and 60. If the predicted shares are respectively 15, 20, and 65, the MAE is (|10-15|+|30-20|+|60-65|)/3 = 6.67. For reporting purposes, we scale the MAE results as a percentage of test/retest reliability. For this study, the test/retest reliability was 3.5. A result of 113% means the predictions were 13% worse than test/retest reliability.

Table 2
HB Draws Performance

Simulation Method	**Error Relative to Test/Retest Reliability**
First Choice	113%
Share of Preference (tuned)	109%

Table 3
HB Point Estimates Performance

Simulation Method	**Error Relative to Test/Retest Reliability**
First Choice	116%
Share of Preference (tuned)	110%
RFC (tuned, E_A only)	108%
RFC (tuned, E_A and E_P)	107%

For simulations using HB draws, we used 50,000 total draws and used both First Choice and Share of Preference models (see Table 2). The First Choice model resulted in an MAE of 3.95, 13% less accurate overall than test/retest reliability. The First Choice shares were a bit too extreme. We found that we could improve matters by using a Share of Preference model. The Share of Preference model, after tuning the exponent, resulted in an MAE of 3.80—9% worse than test/retest reliability.

We then turned to simulations using HB point estimates for the First Choice and Share of Preference models (see Table 3). These were 16% and 10% worse (respectively) than the test/retest reliability. These results were slightly worse than the HB Draw results.

How well did the RFC model predict shares using the point estimates? We used 50,000 replications across respondents. Adding only attribute error, we achieved results of 108%. When product error was added and the relative contribution of attribute and product error tuned, we achieved results of 107%. These predictions are slightly better than the simulations using the HB draws.

The most important finding is that using HB draws in simulations is not better than using RFC on point estimates, despite RFC's simplifying assumptions regarding the error distributions. By using RFC with point estimates, one also avoids having to deal with very large draw files.

> ***Main Point #4: RFC has two important advantages: it produces better predictions than using HB draws, and it avoids dealing with enormous data files.***

There are two interesting effects that may help explain why using RFC with point estimates is more accurate than using draws in simulations for our data set: a *reverse number of levels* effect and an *excluded levels* effect.

"Reverse" Number of Levels Effect

The Number of Levels Effect (NOL) is well documented and prevalent in varying degrees in all types of conjoint analysis studies. The NOL effect as described in the literature is as follows: holding the range of variation constant, if an attribute is defined on more rather than fewer levels, it tends to get more importance, as measured by the range of part-worths for that attribute.

Many researchers have tried to prove an algorithmic explanation to the NOL effect using synthetic, computer-generated data. With the exception of rankings-based card-sort conjoint or pairwise matrices, we are unaware of a previous synthetic data set that has successfully demonstrated a NOL effect. Our research finds when using HB draws a consistent NOL effect using computer-generated data. But the consistent finding is for a "reverse NOL" effect. We use the term "reverse" because it works in the opposite way that we are used to thinking about the NOL effect. Rather than attributes with more levels being biased toward more importance (*ceteris paribus*), those attributes with more levels have *less* importance.

Most treatments of the NOL effect have focused on a measure of importance equal to the range in part-worths for each attribute divided by the sum of the ranges of the part-worths across all attributes. In contrast, this research focuses on the impact an attribute has on what most practitioners use conjoint data for: simulated shares of preference.

HB results in multiple replicates (draws) for each respondent. From those draws, one can estimate within-respondent variances for individual part-worths. It was while studying these

variances that we noticed that attributes with more levels tended to have larger variances around their part-worths and attributes with fewer levels had smaller variances. To illustrate this, we generated a synthetic CBC data set with 6 attributes with known utilities ranging from 1 to 0 within each attribute (equal importances). Half of the attributes had 2 levels (part-worths of 1, 0) and the other half had 4 levels (part-worths of 1, 0.66, 0.33, 0). Three hundred simulated respondents completed 20 tasks with 3 concepts each. Random heterogeneity (between respondents) was added to the part-worths. Five-hundred draws were saved for each respondent. The average part-worths and error variances are presented in Table 4:

Table 4

Part-Worths and Variances
for Synthetic Data Set 1

Attribute	Level#	Avg. Part-Worth	Avg. Within-Person Variance
1	1	2.44	2.86
	2	0.94	3.11
	3	-1.03	3.63
	4	-2.35	3.82
2	1	2.64	1.45
	2	-2.64	1.45
3	1	2.69	3.00
	2	1.23	3.00
	3	-1.12	3.35
	4	-2.79	3.69
4	1	2.97	1.49
	2	-2.97	1.49
5	1	3.16	3.06
	2	0.57	2.69
	3	-0.95	3.38
	4	-2.78	3.33
6	1	2.53	1.44
	2	-2.53	1.44

Importance of attributes 1+3+5: 49.9%
Importance of attributes 2+4+6: 50.1%

Even though the importance of the 4-level attributes (as computed from the aggregate part-worths above) was the same as the 2-level attributes (within 2/10 of a percentage point), the within-respondent variance is much greater around the part-worth estimates than for the 2-level attributes.

If the variances of part-worths are influenced by the number of levels, and the variance of part-worths is directly related to the “scale” of the predictive market choice model, it follows that these differences might lead to systematic biases for simulated shares of preference. To test this hypothesis, we conducted sensitivity simulations.

Shares of choice (First Choice rule) were simulated for each of the (500 draws)(300 respondents), for a total of 150,000 cases. Our approach was one of sensitivity analysis, to test the maximum impact of each attribute on choice versus a constant alternative (with utility of 0, representing an average desirability). For example, starting with attribute one, one enters a product concept made up of levels 1 through 4 (holding all other attributes constant) in four separate simulation steps.

We'll define the "simulated importance" of an attribute as the maximum range of share impact from sensitivity simulations. For example, the shares of choice for attribute one (versus a constant alternative) at each of its four levels was: 0.80, 0.61, 0.36 and 0.22, for a simulated importance of 0.80 - 0.22 = 0.58.

The simulated importances for all attributes are shown in Table 5:

Table 5

Simulated Importances for
Synthetic Data Set 1

Attribute	#levels	Simulated Importance
1	4	0.58
2	2	0.74
3	4	0.64
4	2	0.82
5	4	0.68
6	2	0.75

Avg. Simulated Importance for 4-level Attributes:0.63
Avg. Simulated Importance for 2-level Attributes:0.77

This example demonstrates a consistent NOL effect. The two-level attributes on average have 22% *more* impact in simulations than the four level attributes.

What Causes the "Reverse" NOL Effect?

The variances among HB estimates reflect our uncertainty about parameter values. Consider a balanced design that has some attributes with two levels and others with four. The ratio of observations to parameters to be estimated (within each attribute) is much greater for attributes with fewer levels relative to attributes with more. For the attributes with two levels, its levels occur twice as often within the design relative to attributes with four levels. Therefore, there is more information available to stabilize the part-worths for the two-level attributes.

Which is Bigger: "Reverse" or "Regular" NOL Effect?

The practical question is how big are these effects? Can conducting simulations with HB draws significantly reverse the negative consequences of the usual NOL effect? If it did, would that be a prudent action?

Dick Wittink is the most-published researcher with regard to NOL. To paraphrase his findings over the years, NOL is a significant problem for traditional full-profile conjoint and choice-based conjoint and much less of a problem for ACA.

In the 1997 Sawtooth Software proceedings, Wittink published findings for an experimental study among human respondents rather than "computer" respondents (Wittink 1997). A split-sample study was conducted in which the range of variation was held constant for attributes, but some respondents saw more levels than others did. Two cells in the design (B and C) had the following numbers of levels per attribute (four in total) and resulting importances (as computed by examining the ranges of the aggregate part-worths):

Table 6

(B) Number of Levels	Importance	(C) Number of Levels	Importance
3	25%	2	16%
3	16%	4	24%
3	32%	2	29%
3	28%	4	31%

The percentage gain in importance when increasing the number of levels from 3 to 4 is 1-[(24+31)/(16+28)] = +25%. The net loss in importance when decreasing the number of levels from 3 to 2 is 1-[(16+29)/(25+32)] = -21%. The relative gain for 4-level attributes relative to 2-level attributes is then (1+0.25) / (1-0.21) = +58%.

In the 1999 Sawtooth Software Proceedings, Wittink reported findings from a study by (Shifferstein *et al*. 1998) for another full-profile experiment. Again, human respondents were randomly divided into different groups receiving different versions of the conjoint design. Between cells A and B, the range of variation was held constant, but the number of levels used to describe the attributes differed.

Table 7

(A) Number of Levels	Importance	(B) Number of Levels	Importance
4	0.21	2	0.13
2	0.17	4	0.26
2	0.07	2	0.07

For attribute 1 (the first row), increasing the number of levels from 2 to 4 resulted in a 1-(0.21/0.13) = 62% increase in importance. For attribute two, the same doubling in levels resulted in a 53% increase in importance. We should note, however, that the changes in importance for these two attributes are not independent. From version A to version B, losses in importance (by reducing levels from 4 to 2) for attribute 1 are enhanced by gains in importance (by increasing levels from 2 to 4) for attribute 2. So the net gain in importance (*ceteris paribus*)

one should expect from doubling the number of attributes from 2 to 4 for this data set is something less than either 53% or 62%.

Summarizing the findings from these two full-profile studies (and taking some liberties with the data), doubling the number of levels from 2 to 4 levels (but holding the range of variation for an attribute constant) results in roughly a 50% artificial increase in importance (measured by examining ranges of part-worths).

We noted above that using HB draws in simulations resulted in a reverse NOL effect of about 22% for 4-level relative to 2-level attributes. Again, applying loose math, the "usual" NOL effect is about 2 to 3 times as strong as the reverse NOL effect detected earlier. We might conclude that if we simulated results using HB draws for the two full-profile data sets above, we could cut the NOL effect by about one-third to one-half.

"Reverse" NOL Effect: Good News or Bad News?

If the analyst accepts the usual NOL effect as bad news, anything that counters that effect should be considered good. A counter argument (Wittink 1999b) states that *if* the psychological explanation to NOL holds, there is also likely a NOL effect in the real world for attributes that naturally have different numbers of levels to define available products. This would suggest that our designs should reflect the natural number of level differences and our results should reflect a NOL effect in the part-worths consistent with real world preferences. If methods such as the use of HB draws in simulations consistently reduce that true effect, this then would *not* be a welcomed outcome.

Indeed the argument surrounding NOL is complex. In any case, we aren't comfortable with the temptation to simulate shares from draws to reduce a NOL bias that results from an unbalanced design. Two wrongs don't make a right. We'd prefer to see researchers take appropriate steps to minimize the NOL problem and then simulate shares using RFC and point estimates. Using RFC with point estimates does not reflect the strong reverse NOL bias displayed by HB draws simulations.

"Excluded Level" Effect

After examining a number of HB runs from artificial data sets with known utilities, we noticed that the variance of the last level of each attribute tended to be greater than the variance of the other levels. The reader may notice that the variances for the attribute part-worths presented in Table 4 hint at this effect. It turns out that the greater the number of levels, the more pronounced the difference in variances becomes.

We generated a synthetic data set with two attributes at 12 levels each, with known utilities of zero for all levels (random responses).

Table 8

Attribute	Level#	Within-Respondent Variance
1	1	0.383
	2	0.364
	3	0.417
	4	0.416
	5	0.470
	6	0.385
	7	0.404
	8	0.359
	9	0.359
	10	0.374
	11	0.404
	12	0.856
2	1	0.407
	2	0.371
	3	0.350
	4	0.341
	5	0.430
	6	0.309
	7	0.372
	8	0.427
	9	0.327
	10	0.428
	11	0.315
	12	0.848

Though the expected variances should be equal, the variance of the twelfth level of each attribute is more than double the size of the other levels. Recall that the more variance added to the utility for product concepts, the "flatter" the share of preference in simulations. Therefore, the last levels of each attribute bias the shares for products in which they are included, making them tend toward 50%.

This *excluded level effect* is an artifact resulting from the effects-coding procedure used in coding the independent variable matrix. Effects-coding constrains part-worths to be zero-centered. The last level is "excluded" from the design and solved later as negative the sum of the effects of the other levels within the same attribute.

We are indebted to Rich Johnson for providing this explanation regarding why the excluded level has a much higher variance:

> "The interesting curiosity is that the final (excluded) level has variance much greater than that of the other (included) levels, and the discrepancy increases as the number of levels in the attribute. Of course the reason for this is that the variance of a sum is equal to the sum of the variances and covariances. If the covariances among levels were zero, then the variance of the excluded level would be (n -1) times as large as the included levels, where n is the number of attributes. Since the covariances are for the most part negative, the actual effect is smaller than that, but still sizeable.
>
> "One doesn't see that effect with logit or other aggregate methods because in that case the

expansion is done on point estimates, which have small variances. But when we do it on individual draws, the effect looms large.

"As one might expect, a similar but opposite thing occurs with dummy-variable coding. In that case the excluded level is estimated by taking the negative of the mean of the remaining levels, so one would expect its variance to be smaller. That turns out to be the case. There appears to be no way around this problem, which may limit the usefulness of HB draws in first choice simulations."

The *excluded level effect* does not exist for two-level attributes, and is very minor for attributes with only a few more levels than that. For attributes with many levels, it could conceivably lead to significant biases in market simulations when using the HB draws.

Main Point #5: Two reasons why RFC may work better are that HB draws have a reverse NOL effect and an excluded level effect.

Summary and Conclusions

We have reviewed the two most widely used methods for simulating choices from conjoint or choice part-worths, namely the First Choice and Share of Preference (logit) rules. The First Choice model is immune to IIA difficulties, but is often too steep and is not tunable. The logit rule is tunable, but suffers from IIA. Randomized First Choice (RFC) combines benefits of both models and can improve the predictive accuracy of market simulations.

Like RFC simulations, hierarchical Bayes draws reflect uncertainty about the point estimates for part-worths. But, within the unit of analysis, RFC assumes a covariance matrix for part-worths with equal variances along the diagonal and zeroes in the off-diagonal elements. HB makes neither of these assumptions: the variances of the part-worths can differ, and covariances are not assumed to be zero. We compared the predictive accuracy of RFC simulations on point estimates versus conducting simulations using HB draws. The RFC simulations were slightly more accurate for our data set, and they avoided having to use the huge draw files.

We also demonstrated that using HB draws in simulations is subject to two biases: a *reverse number of levels* effect, and *an excluded level* effect. These biases have the potential to significantly degrade the predictive accuracy of market simulations.

A number of sophisticated approaches have been suggested for circumventing IIA and improving the predictive validity of market simulations. These techniques have included Mother Logit, Multinomial Probit and Nested Logit. We have not attempted to test or expound on those techniques here. In our opinion, for "in-the-trenches" practical research, a well-tuned RFC model operating on well-developed individual-level point estimates from HB estimation is hard to beat.

References

Huber, Joel, Orme, Bryan K. and Richard Miller (1999), "Dealing with Product Similarity in Conjoint Simulations," Sawtooth Software Conference Proceedings, pp 253-66.

McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka (ed.) Frontiers in Econometrics. New York: Academic Press.

Orme, Bryan (1998), "Reducing the IIA Problem with a Randomized First Choice Model," Working Paper, Sawtooth Software, Sequim, WA.

Sawtooth Software (1999), "The CBC System, v2.0," Sequim, WA.

Sawtooth Software (1999), "The CBC/HB Technical Paper," http://www.sawtoothsoftware.com/TechPap.htm.

Shifferstein, Hendrik N. J., Peter W. J. Verlegh and Dick R. Wittink (1998), "Range and Number-of-Levels Effects in Derived and Stated Attribute Importances," Working Paper.

Wittink, Dick R. (1997), "Solving the Number-of-Attribute-Levels Problem in Conjoint Analysis," Sawtooth Software Conference Proceedings, pp 227-40.

Wittink, Dick R. (1999a), "A Comparison of Alternative Solutions to the Number-of-Levels Effect," Sawtooth Software Conference Proceedings, pp 269-84.

Wittink, Dick R. (1999b) "Comment on McCullough," Sawtooth Software Conference Proceedings, pp 117-21.