17 Dealing with Product Similarity in Conjoint Simulations[1]

Joel Huber, Bryan Orme and Richard Miller

17.1 The Value of Choice Simulators

One of the reasons conjoint analysis has been so popular as a management decision tool has been the availability of a choice simulator. These simulators often arrive in the form of a software or spreadsheet program accompanying the output of a conjoint study. These simulators enable managers to perform 'what if' questions about their market - estimating market shares under various assumptions about competition and their own offerings. As examples, simulators can predict the market share of a new offering; they can estimate the direct and cross elasticity of price changes within a market, or they can form the logical guide to strategic simulations that anticipate short- and long-term competitive responses (Green and Krieger 1988).

Choice simulators have four stages. The first stage estimates a preference model for each individual or homogeneous segment in the survey. The second stage defines the characteristics of the competitors whose shares need to be estimated. The third stage applies the preference model to the competitive set to arrive at choice probabilities for each alternative and each segment or respondent. The final stage aggregates these probabilities across segments or individuals to predict choice shares for the market.

We pay the most attention to the third stage-estimating choice probabilities for each individual or segment. We explore the value of adjusting individual choice probabilities with two kinds of variability, each of which has a simple intuitive meaning. The first kind, product variability, occurs when a consumer simply chooses a different alternative on different choice occasions, typically through inconsistency in evaluating the alternatives. The second kind, attribute variability, occurs when a consumer is inconsistent in the relative weights or part worths applied to the attributes. As an example of this second kind of variability, consider a consumer who notices the nutrition label on breads in one shopping trip but is price sensitive in other trips. While most simulators do not distinguish between these two forms of variability, we will show that they differ strongly in their treatment of similarity. Attribute variability preserves appropriate similarity relationships among alternatives while product variability clouds them. However, attribute variability by itself allows for no residual error in choice once the part

---

[1] Originally presented at the Sawtooth Software Conference, February 2, 1999 and updated for this volume in 2006.

worth values have been simulated. Thus, to appropriately model individual choice it is necessary to include both sources of variability.

We present Randomized First Choice as a general way to „tune" conjoint simulators to market behavior. Conceptually, Randomized First Choice begins with the assumption of no variability - the highest utility alternative in the set is chosen all the time. Then it adds back levels of attribute and alternative variablity that best match choice shares in the environment. This process allows sufficient flexibility to approximate quite complex market behavior.

Mathematically, Randomized First Choice adds variation in the attribute values in addition to variation in the final product valuation. It begins with a random utility model with variability components on both the coefficients and the residual error:

$$(1) \qquad U_i = X_i (\beta + E_A) + E_P$$

where:

U_i = Utility of product i for an individual or homogeneous segment at a moment in time
X_i = Row vector of attribute scores for alternative i
β = Vector of part worths
E_A = Variability added to the part worths (same for all alternatives)
E_P = Variability added to product *i* (unique for each alternative)

In the simulator, the probability of choosing alternative *i* in choice set S is the probability that its randomized utility is the greatest in the set, or:

$$(2) \qquad Pr(i|S) = Pr(U_i \geq U_j \text{ all } j \in S).$$

Equation 2 is estimated by using a simulator to draw U_i's from equation 1 and then simply enumerating the probabilities. To stabilize shares, group or individual choices are simulated numerous times.

Those familiar with logit will recognize that E_P is simply the error level in the logit model. The typical adjustment for scale in the logit model is mathematically equivalent to adjusting the variance of a Gumbel-distributed E_P in RFC simulations. The E_A term then reflects taste variation as has been found in models by Hausman and Wise (1978) and in work in mixed logit by Revelt and Train (1998).

The purpose of this paper is to provide an understanding of why including attribute variability is superior to just including product variability. The quick answer is that attribute variability is needed to account for expected similarity relationships whereas adding product variability clouds those relationships. The next section begins by detailing the desirable properties of any choice simulator. Then follows an experiment that demonstrates the effectiveness of adding attribute and product variability, particularly when applied to aggregate and latent class seg-

ments, but also for individual choice models generated by hierarchical Bayes and Sawtooth Software's ICE (Individual Choice Estimation).

17.2 Three Critical Properties of Market Simulators

Market simulators need three properties if they are to reflect the complexity of market behavior. First, the individual- or segment-level model must display differential impact - where the impact of a marketing action occurs as an alternative in a competitive set reaches the threshold for choice. Second, the model needs to exhibit differential substitution, a property where new alternatives take disproportionate share from similar competitors. Finally, the simulator must display differential enhancement, the idea that very similar pairs can produce disproportionately severe choice probabilities. Each of these is detailed below.

Differential Impact is a central requirement of an effective choice simulator. It reflects the property that the impact of a marketing action depends on the extent that the alternative is near the purchase threshold. This point of maximum sensitivity occurs when the value of an alternative is close to that of the most valued alternatives in the set - when the customer is on the cusp with respect to choosing the company's offering. At that time, an incremental feature or benefit is most likely to win the business.

The differential impact implicit in a threshold model can best be understood by examining three cases reflecting different kinds of thresholds. First we present the linear probability model which importantly defines the case of no threshold. Then we examine the other extreme, that of a first choice model, which has the most extreme step-like threshold. Finally we consider the standard choice models (logit, probit) whose threshold has been softened by the addition of variability.

If probability is a linear function of utility, then improving an attribute has the same effect on choice share regardless of how well it is liked. There are many problems with this linear probability model, the worst of which is a lack of differential impact. Under a linear probability model adding, say, an internal fax modem has the same share impact regardless of whether it is added to a high- or low-end computer. By contrast, a threshold choice model specifies that the benefit from adding the modem mainly affects those consumers who are likely to change their behavior. This makes good sense - adding the feature does not affect a person who would have bought the brand anyway, nor does it affect customers who would never consider it. Managerially, the differential impact brought about by a threshold model has the benefit of focusing managerial attention on the critical marginal customer, and thereby avoids expensive actions that are unlikely to alter market behavior.

The first-choice model offers an extreme contrast to the linear model. The first choice model is mathematically equivalent to Equation 1 with no variability ($var(E_P) = var(E_A) = 0$). In the first choice simulation, share of an alternative is zero until its value is greater than others in the set. Once its value exceeds that threshold, however, it receives 100%. The problem with the first choice model is that it is patently false. We know that people do not make choices without vari-

ability. In studies of experimental choices, given the same choice set (3-4 alternatives, 4-5 attributes) respondents choose a different alternative about 20% of the time. In our study, respondents chose a different alternative in the repeated task 19% of the time. One of the paradoxes we hope to resolve in this paper is why the first choice model operating on individual-level part worths works so well despite its counter-factual premise.

Standard logit and probit models reflect a compromise between the first-choice and linear model. Instead of the severe step function characteristic of the first choice model, the variablity implicit in these models moderates the step into a smooth s-shape or sigmoid function. As shown in Equations 1 and 2, these models are identical to first-choice models with variability added. For logit, E_P has a Gumbel, while for Probit, it has a Normal distribution. It is important to note, however, that these models are, to use a technical phrase, linear-in-the-parameters. Thus the *utility* of an item generally increases the same amount with a given improvement, however, the *probability of purchase* follows a threshold model.

A little-understood benefit of a threshold model is that it can reflect complex patterns of interactions between, say, a feature and a particular brand simply through the simulation process. An interaction term specifies that a particular feature has a differential impact on particular brands. While these interaction terms can be reflected in the utility function, we propose that many interactions can be better represented as arising from the aggregation of heterogeneous customers each following a threshold model. For example, consider a warranty x price interaction indicating that a warranty is more valuable for low- over high-priced appliances. The same effect could also emerge in a simulation of respondents under a threshold rule. Suppose there are two segments, one valuing low price and the other desiring high quality. Adding a warranty to the low-priced brand might not be sufficient to raise it past the purchase threshold of those desiring high quality. By contrast, the warranty pushes the alternative past the threshold of those desiring low prices. When these two segments are aggregated it appears that the warranty mainly helps the low priced brand and thus appears to justify an interaction term in the utility function. However, the same behavior can be reflected in a simulator with a threshold model. The heterogeneity account has the further advantage of being more managerial actionable than the curve-fitting exercise of the cross term.

The greatest difficulty with interaction terms is that their numbers can grow uncontrollably large. Above we illustrated an example of price tiers, but there can be many others. Consider combinations of brand tiers where customers are simply not interested in certain brands; size tiers where a large size never passes the threshold for certain segments, and feature tiers, where certain groups are only interested in certain features. Modeling these with interaction terms in the utility function is both complicated and can lead to problems with overfitting or misspecification. The beauty of a simulator operating on segmented or individual models is that it can approximate this behavior in the context of a simple main-effects additive model (e.g., see as Orme and Heft).

To summarize, differential impact is critical if we believe that impact on choice of, say, a new feature of a brand depends on values of the brands against

which it competes. The threshold model within a random utility formulation focuses managerial attention on those alternatives that are on the cusp, and in that way places less emphasis on alternatives that are already chosen, or would never be. Further, applying the threshold model at the level of the individual or homogeneous segment confers the additional benefit of isolating the differential impact appropriately within each.

Differential Substitution is the second property critical to an effective choice simulator. Its intuition follows from the idea that a new offering takes share disproportionately from similar ones. Differential substitution is particularly important because the dominant choice model, aggregate logit displays *no* differential substitution. The logit assumption of proportionality implies that a new offering that gets, say, 20% of a market will take from each competitor in proportion to its initial share. Thus a brand with an initial 40% share loses 8 percentage points (40% x .2) and one with 10% share loses 2 percentage points (10% x .2). Proportionality provides a naive estimate of substitution effects and can result in managerially distorted projections where there are large differences in the degree of similarity among brands in the market. For example, a product line extension can be expected to take proportionately most share from its sibling brands. Managers recognize this problem. Successful companies manage their portfolios with new brands that are strategically designed to maximize share taken from competitors and minimize internal share losses. By contrast, proportionality glosses over such strategically important distinctions. Ignoring differential substitution could lead to the managerial nightmare of numerous line extensions whose cost to current brands is regularly underestimated.

An extreme, if instructive, example of differential substitution is the presence of a duplicate offering in the choice set. Economic theory often posits that a duplicate offering should take half the share of its twin, but none from its competitor. However, in practice this expectation is rarely met. If some consumers randomly pick a brand without deleting duplicates, then having a duplicate could increase total choice share. Indeed, the fight for shelf space is directed at capturing that random choice in the marketplace. To the extent that a duplicate brand increases the total share for that brand, we label the increase in total share from a duplicate *share inflation*. Clearly some share inflation is needed, but it is unclear how much. In the empirical test we measure the extent to which simulators reflect differential enhancement by how well they correctly predict the combined share of near substitutes in the holdout choice sets.

Differential enhancement is the third property needed by choice simulators. It specifies a second, but less commonly recognized way product similarity affects choices. Under differential enhancement, pairs of highly similar alternatives display more severe choice differences. Psychologically, this phenomenon derives from the idea that similar alternatives are often easier to compare than dissimilar ones. Consider the choice between French Roast coffee, Jamaican Blend coffee and English Breakfast tea. A change in the relative freshness of the coffees can be expected to enhance the relative share of the fresher coffee, while having relatively little impact on the proportion choosing tea.

In its extreme form, differential enhancement arises where one offering *dominates* another in the choice set. Rational economic theory typically posits that the dominated alternative receives no share, while the shares of the other brands are unaffected. Market behavior is rarely as neat. There are few purely dominated alternatives in the market. Even finding two otherwise identical cans of peas in the supermarket can lead to suspicion that the lower priced one is older. Determining dominance requires work that consumers may be unwilling or unable to perform. For that reason, manufacturers intentionally create differences between offerings (new line, different price, channel), so that dominance, or near dominance is less apparent. From a modeling perspective, the important point is that any choice simulator needs to allow both for dominance to produce cases of extreme probability differences and to allow consumers to be fallible in their ability to recognize that dominance.

The modeling implications of differential enhancement parallel those for differential substitution. The standard logit or probit models assume that the relative shares of any pair of alternatives only depend on their values, not on their relative similarity. Referring to a classic example, if trips to Paris and to London are equally valued, then a logit model predicts that adding a second trip to Paris with a one-dollar discount will result in one-third shares for the three alternatives. There are numerous ways researchers have attempted to solve this problem, from nested logit to correlated error terms within probit. Within the Sawtooth Software family Model 3 penalizes items that share attribute levels with other alternatives in the choice set. We will show that a simple first choice simulation with suitable variability added to both attributes and alternatives provides a robust way to mirror these complex market realities.

17.3 A Market Study to Validate Choice Simulators

As we approached the task of comparing the ability of different choice simulators to deal with varying degrees of alternative similarity, it became apparent that choice sets typically used for choice experiments would not work discriminate between models. For the sake of efficiency, most choice experiments feature alternatives where the numbers of levels differing among pairs of alternatives are relatively constant. For example, it would not typically make sense to include a near alternative twice since its inclusion adds so little additional information. In this study we deliberately add alternatives which are duplicates or near duplicates to be able to test the ability of various simulators to appropriately handle these difficult choices.

Three hundred ninety-eight respondents completed computerized surveys in a mall intercept conducted by Consumer Pulse, Inc. The survey involved preference for mid-sized televisions and was programmed using Sawtooth Software's Ci3 and CBC systems. Respondents over 18 who owned a television or were considering purchasing a mid-sized television set in the next 12 months qualified for the survey. The first part of the interview focused on attribute definitions (described in terms of benefits) for the six attributes included in the design. The main part of the

survey involved 27 choices among televisions they might purchase. Each choice involved five televisions described with six attributes: brand name (3 levels), screen size (3 levels), picture-in-picture (available, not), channel blockout (available, not) and price (4 levels). Table 1 gives an example of a choice set that illustrates the levels. We gave respondents a $4.00 incentive to complete the survey, and urged them to respond carefully.

Table 1: Example of a Holdout Choice Set

25" JVC, Stereo, Picture in Picture, No Blockout, $350	26" RCA, Surround Sound, Picture in Picture, Blockout, $400	25" JVC, Monaural, No Picture in Picture, No Blockout $300
27" Sony, Surround Sound, No Picture in Picture, No Blockout $450	25" JVC, Stereo, Picture in Picture, No Blockout, $350	

Preliminary data from a small pre-test suggested that respondents were not giving sufficient effort to answer consistently. In an attempt to improve the quality of the data, we revised the survey. We told them that the computer would „learn" from their previous answers and know if they were answering carefully or not. The „computer" would reward them with an extra $1.00 at the end of the survey if they had „taken their time and done their task well." (We displayed a password for them to tell the attendant.) In terms of programming the survey logic, we rewarded them based on a combination of elapsed time for a particular section of the survey and test-retest reliability for a repeated holdout task. Though it is difficult to prove (given the small sample size of the pretest), we believe the revision resulted in cleaner data. Nearly two-thirds of the 398 respondents received the extra dollar. We discarded 46 respondents based on response times to choice tasks that were unusually low, leaving 352 for analysis.

The first 18 choice tasks were CBC randomized choice sets that did not include a „None" option. After completing the CBC tasks, respondents were shown an additional nine holdout choice tasks, again including five alternatives. The holdout tasks were different in two respects. First, to test the market share predictions of the different simulators, it was critical to have target sets for which market shares could be estimated. Respondents were randomly divided into four groups with approximately 90 in each group that would receive the same nine holdout choice tasks. Additionally, we designed the holdout choices to have some ex-

tremely similar alternatives. Four of the five alternatives in the holdout tasks were carefully designed to have approximate utility and level balance (Huber and Zwerina 1996). However, the fifth alternative duplicated another alternative in the set, or duplicated all attributes except the two judged least important in a pretest. To provide an estimate of test-retest reliability, each respondent evaluated two choice sets that were perfect replicates. Across respondents, the computer randomized both choice set and product concept order.

17.4 The Contenders

We analyzed the CBC data using four base methods for estimating respondent part worth utilities: Aggregate Logit, Latent Class, Sawtooth Software's ICE (Individual Choice Estimation) and Hierarchical Bayes (courtesy of Neeraj Arora, Virginia Tech). There is logic behind picking these four methods. Aggregate logit is important in that it reflects what happens when all respondents are pooled into one choice model. By contrast, latent class analysis seeks sets of latent segments (we used an eight-group solution) whose part worths best reflect the heterogeneity underlying the choices (Kamakura and Russell 1989; Chintagunta, Jain and Vilcassim 1991; DeSarbo, Ramaswamy and Cohen 1995). ICE then takes these segments and builds a logit model that predicts each individual's choices as a function of these segments (Johnson 1997). It thereby is able to estimate a utility function for each person. Hierarchical Bayes assumes respondents are random draws from a distribution of part worth utilities with a specific mean and variance. It produces a posterior estimate of each individual's part worths reflecting the heterogeneous prior conditioned by the particular choices each individual makes (Lenk, DeSarbo, Green and Young 1996; Arora, Allenby and Ginter 1998). Both ICE and hierarchical Bayes reflect current attempts to generate each individual's utility functions from choice data, while latent class and aggregate logit typify popular ways to deal with markets as groups.

For each of these base models we examine the impact of adding three levels of variability within the Randomized First Choice framework. The initial condition is the first choice rule that assumes respondents choose the highest valued alternative in a choice set with certainty. The second condition adds the level of product variability that best predicts holdout choice shares. This latter condition is identical to adjusting the scale under the logit rule to best predict these shares. The third condition tunes both product and attribute variability to best predict the holdout choice shares. The mechanism of the tuning process is simple but tedious: we use a grid search of different levels of each type of variability until we find those that minimize the mean absolute error in predicting holdout choice shares.

17.5 Results

We examine the ability of different simulators to handle product similarity from different perspectives. First, we measure deviations from predicted and actual

share for the duplicates and near-duplicates that were included in the holdout choice sets. This focus enables us to uncover ways the various models appropriately account for differential substitution and differential enhancement. Then we broaden our perspective to consider the overall fit of the models - how well the models predict choice shares for all items in the choice set.

Differential substitution requires that similar items take disproportionate share from each other. Thus, our near and perfect substitutes should cannibalize share from each other. For example, if an alternative would receive 20% share individually, the joint share of the two alternatives should be only marginally more than 20%, since the new one takes most of its share from its twin. A first choice simulator, with its assumption of zero variability puts the joint share at exactly 20%, but in the marketplace this combined share is likely to be somewhat higher. Put differently, due to fundamental noise in the consumer choice processes we can expect some share inflation.

Table 2 gives predicted combined share of the near and perfect substitutes divided by the actual share. Thus, a value of 100% means that the degree of differential substitution reflected in the holdout choices was estimated perfectly. Notice that the first choice rule underestimates the joint share of the near substitutes by about 10%, indicating that the first choice rule of no variability is too severe. The next column shows the result of adding the level of product variability that best predicts the holdouts. In this case, adding that variability seriously overestimates the share inflation for the near substitutes, in effect, assuming too much variability. The third column then adjusts both product and attribute variability to optimally predict choice shares. By allowing some attribute variability to substitute for product variability, we are able to more closely track actual differential substitution in this data set for all models except ICE.

It is also instructive to compare the rows representing the four core models. The two aggregate models, logit and latent class, suffer most from overestimation of share inflation under product variability. However, when both forms of variability are combined, they do remarkably well. The two individual models appear both less sensitive to the addition of variation and less in need of it. We will discuss the implications of this phenomenon after the other results from the study are presented.

Differential enhancement occurs when a given quality difference results in a greater share difference between highly similar pairs. We examine the share difference between the alternative with higher expected share and its near duplicate. Table 3 gives the model's prediction of this difference as a percent of the actual difference. Once again a score of 100% indicates perfect level of differential enhancement relative to the actual choices.

The two aggregate models with product variability only are the least effective in representing the differential enhancement reflected in the holdout choices. In contrast, the first choice rule applied to the individual level models performs very well in this case. In all cases, adding the optimal level of product variability tends to understate desired levels of differential enhancement. Optimizing both kinds of variability has a small incremental benefit but results in predictions that still underestimate the appropriate level of differential enhancement.

Table 2: Differential Substitution: Predicted Combined Share of Near Substitutes As Percent of Actual Share

	First Choice Rule	Product Variability	+Attribute Variability
Aggregate Logit	N/A	139%	108%
Latent Class	N/A	119%	105%
Hierarchical Bayes	91%	117%	104%
ICE	89%	101%	94%

Table 3: Differential Enhancement: Predicted Difference between Similar Alternatives As Percent of Actual Differences

	First Choice Rule	Product Variability	+Attribute Variability
Aggregate Logit	N/A	63%	73%
Latent Class	N/A	71%	74%
Hierarchical Bayes	100%	73%	77%
ICE	90%	77%	79%

It needs to be emphasized that these measures of differential substitution and enhancement only relate to the shares of near substitutes. By contrast, the optimization to choice shares counts all five alternatives, not just the two most similar ones. The overestimation of differential substitution shown in the last column of Table 2 and the underestimation of differential enhancement in the last column of Table 3 could have been improved by decreasing the level of product variability, but overall fit would have suffered. An interesting implication of this result is that the actual variability around judgments relating to the share sums and share differ-

ences of these near substitutes may be smaller than for alternatives generally. An interesting path for future research involves allowing variability to change as a function of similarity of an alternative within each set.

Table 4: Relative Error: Mean Absolute Error Predicting Market Share As Percent of Test-Retest

	First Choice Rule	Product Variability	+Attribute Variability
Aggregate Logit	N/A	151%	112%
Latent Class	N/A	117%	105%
Hierarchical Bayes	125%	110%	107%
ICE	112%	106%	106%

Relative error measures the degree that the different simulators predict the market shares across all alternatives in the holdout tasks for the study. Table 4 shows mean absolute error (MAE) predicting holdout stimuli as a percent of the test-retest MAE for repeated choice sets. For example, the 151% for aggregate logit indicates that adding product variability only results in an error that is about one and one-half times as great as for the choice replication. Adding attribute variability helps all models, but the greatest gains occur for the aggregate models.

Table 4 offers several surprises. The first surprise is that Randomized First Choice applied to latent class does as well as any of the models. The positive impact of both kinds of variability on latent class makes sense because the original latent class model assumes that there is no heterogeneity within each latent class. By optimizing both product and attribute variability we are able to transform latent class from an elegant but counterfactual model into one that tracks choice shares remarkably well.

The second surprise is that the addition of attribute variability has very little impact on either of the individual level models. For both hierarchical Bayes and ICE the addition of product variability is the major benefit. We believe there is a simple reason for this result. The individual level models are not estimated with perfect accuracy, but have significant variation due to the noise in individual choices and the fact that many parameters are being estimated from relatively few observations. Thus, when estimates from these models are put in a simulator they act as if variability has already been added to the part worths. However, in this

case instead of attribute variability coming from the RFC process, it comes from the inherent variability in the estimation model. This insight then leads to an important conclusion: where variability in the estimation technique is greater than in the market, then the optimal variability to add to the first choice model will be zero (see also Elrod and Kumar 1989).

The final surprise is that Randomized First Choice predictions are quite good regardless of the core estimation method used (except aggregate logit). That is, using RFC produces accuracy that is within 10% of what one would get asking the same question again. Clearly few techniques are going to do much better than that. There simply is not much room for further improvement.

Before concluding, it is important to briefly mention Sawtooth Software's Model 3, a long-available method that accounts for item similarity in a simulation. Model 3 operates by penalizing alternatives that have high numbers of levels in common with other attributes in a choice set. It does so in such a way that adding a perfect duplicate perfectly splits share with its twin when these duplicates share no levels in common with the other alternatives. Model 3 acts like the first choice model in assuming that there is zero share inflation from adding an identical alternative, thereby underestimating the joint share of the two identical alternatives for the holdout choices in our study. Further, Model 3 reflects a relatively simple (and inflexible) rule regarding differential substitution and does not address differential enhancement at all. Since Model 3 is not a theoretically complete model of similarity effects, it did not surprise us that for our study Model 3 was consistently outperformed by RFC. In our view, Sawtooth Software users should replace Model 3 with RFC.

17.6 Summary and Conclusions

The purpose of this paper has been to examine ways to build choice simulators that correctly reflect similarity effects. We began with the introduction of three principles needed for sound conjoint simulations, and in the light of those principles developed Randomized First Choice. RFC provides better choice share predictions by determining the optimal levels of attribute and product variability when generating simulated choices.

The first requirement of effective simulators is that they reflect differential impact. This property permits the simulator to focus managerial attention on those actions that are most likely to impact their customers. In addition, a little-known implication of the threshold model at the level of a segmented (e.g. latent class) or individual model is that it automatically allows for various kinds of price and offering tiers without the necessity of interaction terms. The cost of losing that benefit is best illustrated by the poor performance of the aggregate logit simulation, even with variability added. In simple, main-effects aggregate logit, there is no way the threshold effect can display the action of different segments. Either the homogeneous segments from latent class or individual models are necessary for that benefit.

Effective simulators also need to reflect differential substitution. Our analysis of the combined share of near and perfect substitutes indicates that the first choice model underestimates, while adding product variablity overestimates their combined share. The joint optimizations of both product and attribute variability then permit the estimates of combined share to closely approximate the actual choices. One can tune the appropriate balance of differential substitution/share inflation.

The third requirement of effective simulators is that they demonstrate differential enhancement. We illustrated this requirement by examining the share difference of nearly identical alternatives. The first choice rule overestimates differential enhancement in aggregate models by giving all share to the preferred alternative. By contrast, adding product variability underestimates the predicted share differences. Adjusting both kinds of variability improved this underestimation but did not solve it completely. Since differential enhancement comes in part from a psychological mechanism whereby decisions between similar alternatives are easier, a full solution to this problem may await models that adjust item variability to the difficulty in making the choice.

We demonstrated the benefits of RFC on a study in which the holdout choices included „difficult" alternatives that included near and true duplicates. However, a greater benefit for Sawtooth Software users may come in contexts where it is possible to project to actual market shares. Most markets will have far more complicated similarity structures than our simple problem, resulting from competition among family brands, different sizes, price tiers and subbrands. We believe that RFC with its two kinds of variability will be very useful in tuning the simulator to successfully account for market behavior in such cases.

17.7 References

Arora, N., Allenby, G. and Ginter, J. L. (1998), A Hierarchical Bayes Model of Primary and Secondary Demand, *Marketing Science*, 17, 29-44.

Chintagunta, P., Jain, D. C. and Vilcassim, N. J. (1991), Investigating Heterogeneity in Brand Preferences in Logit Models for Panel Data, *Journal of Marketing Research*, 28, 417-428.

DeSarbo, W. S., Ramaswamy, V. and Cohen, S. H. (1995), Market Segmentation with Choice-Based Conjoint Analysis, *Marketing Letters*, 6, 137-148.

Elrod, T. and Kumar, S. K. (1989), Bias in the First Choice Rule for Predicting Share, *Sawtooth Software Conference Proceedings*.

Green, P. E and Krieger, A. M. (1988), Choice Rules and Sensitivity Analysis in Conjoint Simulators, *Journal of the Academy of Marketing Science*, 16, 114-127.

Hausman, J. and Wise, G. (1978), A Conditional Probit Model for Quantitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences, *Econometrica*, 43, 403-426.

Huber, J. and Zwerina, K. (1996), The Importance of Utility Balance in Efficient Choice Designs, *Journal of Marketing Research*, 23, 307-317.

Johnson, R. M. (1997), ICE: Individual Choice Estimation, *Sawtooth Software Technical Paper*.

Kamakura, W. A. and Russell, G. J. (1989), A Probabilistic Choice Model for Market Segmentation and Elasticity Structure, *Journal of Marketing Research*, 26, 339-390.

Lenk, P. J., DeSarbo, W. S., Green, P. E. and Young, M. R. (1996), Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs *Marketing Science*, 15, 173-191.

Orme, B. K. and Heft, M. (1999), Predicting Actual Sales with CBC: How Capturing Heterogeneity Improves Results, *Sawtooth Software Conference Proceedings*.

Revelt, D. and Train, D. (1998), Mixed Logit with Repeated Choices: Household's Choices of Appliance Efficiency Level, *Review of Economics and Statistics*, forthcoming.

Rossi, P. and Allenby, G. (1993), A Bayesian Approach to Estimating Household Parameters, *Journal of Marketing Research*, 30, 171-182.

Addendum 2006

For this latest edition of *Conjoint Measurement*, the authors asked us to revisit this article and provide new insights and an update.

After publishing this article here and in the *1999 Sawtooth Software Proceedings* we later re-analyzed the data and published the results in *Marketing Research*, winter 2000. After writing the earlier articles, we had recognized that in tuning the two types of error in Randomized First Choice (RFC) to best predict holdouts, we were risking the possibility of overfitting and potentially overstating the benefit of RFC. We addressed this possibility in the 2000 *Marketing Research* article by splitting the sample into two matched replicates. We re-estimated the models within each of the replicates (this time using Sawtooth Software's commercial CBC/HB routine, which had not been available for the earlier work). We tuned the product and attribute errors for RFC on the first half of the sample and applied those error amounts to new respondents in the other half. Only the second group of respondents was used for predicting holdouts. The new overall error (relative to test-retest reliability) is given below, with the previous error rates as published in this article shown in parentheses:

	First Choice Rule	Product Variability	+Attribute Variability
Aggregate Logit	N/A	155% (151%)	121% (112%)
Latent Class	N/A	122% (117%)	113% (105%)
Hierarchical Bayes	116% (125%)	111% (110%)	107% (107%)

We were pleased to see that the essential findings held: RFC offered an improvement over tuning only for scale (product variability only). As before, the benefit was greatest for the aggregate models.

We have now had about eight years experience working with Randomized First Choice simulation models. Sawtooth Software added the capability to its commercial market simulator, even making it the default simulation method. In general, it has worked well. It should not surprise the reader that we have learned a few things: both how to improve RFC and also regarding weaknesses.

Eight years ago, the majority of Sawtooth Software customers were using aggregate models: logit or latent class. RFC clearly provided a benefit in these cases. Lately, the majority of Sawtooth Software customers are using part worths estimated under HB. For these customers, RFC provides modest improvements. Thus, the popularity and effectiveness of HB has in turn reduced the impact that RFC has in our industry.

Some HB users (especially academics) prefer to use the draws within choice simulators rather than the point estimates, as we applied in this research. It could be argued that HB provides more empirically correct draws of random error around point estimates (parameter-specific estimated variances) rather than RFC's simple assumption of uniform error variance across the parameters. In 2000, Orme and Baker compared the use of RFC to HB draws, again in terms of fitting holdout choices. They tuned both HB draws (product error only) and RFC operating on the HB point estimates (both product and attribute error) to best predict holdouts. The relative error rates were 107% and 109% for RFC and HB draws, respectively. The authors concluded that using the huge HB draws file was unnecessary, and RFC's simpler model performed equally well or better.

In the 2004 Sawtooth Software Conference, Allenby *et al.* pointed out that standard HB models can face what they termed "IIA Meltdown" when very many alternatives (such as 84 alternatives in a beverage category or even more alternatives in the automobile category) are in the choice design. Although they proposed a different model from RFC, their findings that the standard HB simulators face greater IIA troubles as the number of alternatives increased suggests that RFC may be even more useful in these cases.

We have also noted a weakness with RFC simulations. The simple RFC model assumes that all attributes involve a correction for product similarity. However, it is not clear that this should be the case. For instance, price represents an attribute for which it isn't clear that product similarity should apply. Some analysts like to derive demand curves via sensitivity analysis within choice simulators. Under RFC, if all products are first aligned on the average price (and the "test" product systematically varied across all price levels), an unwanted "kink" will occur in the demand curves around the point that was artificially chosen as the average price. When the test product is changed from the average to the next higher price point (and all others remain at the average price), it sometimes receives a boost in share due to its becoming less similar that nearly counteracts the penalty from becoming more expensive. One solution to this problem is to apply independent error to the price part worths for all alternatives in the simulation. This gives rise to a more sophisticated RFC simulator, where some attributes involve correlated error (when

product alternatives share the same levels of these attributes) and other attributes involve uncorrelated errors (when product alternatives receive independent random error draws for these attributes, irrespective of shared levels).

There is another opportunity for analysts to improve RFC modeling. Some conjoint/choice designs involve many alternatives. Beverages and automobiles are good examples. Suppose we had conducted a choice study with 200 automobile makes, including trucks, minivans, sedans, and coupes. Further suppose that we had treated the makes (for part worth estimation) as independent levels of a 200-level attribute. However, we know that these 200 makes fall into four clear categories that should reflect increased competition within each category. One could assume a new attribute with four levels (truck, minivan, sedan, and coupe) for which we apply attribute-type error under RFC in choice simulations.

We should also note that as the number of alternatives in the simulated choice scenario increases, the number of draws used in RFC should also be increased. Otherwise, the random error involved with RFC may be uncomfortably large relative to the signal associated with some relatively tiny product shares.

Finally, given the strong performance of ICE (Individual Choice Estimation) for this data set, the reader might wonder why Sawtooth Software's ICE program is not used much any more and essentially has been abandoned by Sawtooth Software. Although it worked quite well for this dataset (and many others), it hasn't been as generally robust as HB. ICE can be problematic with sparse datasets (many parameters to estimate relative to the number of choice tasks at the individual level). HB has been widely embraced by the industry and is more theoretically appealing. Given the speed of computers today, it is also very manageable for most every practical data set, with run times seldom exceeding a few hours.

Additional References:

Allenby, Greg, Jeff Brazell, Tim Gilbride, and Thomas Otter (2004), Avoiding IIA Meltdown: Choice Modeling with Many Alternatives, *Sawtooth Software Conference Proceedings*.

Orme, Bryan and Gary Baker (2000), Comparing Hierarchical Bayes Draws and Randomized First Choice for Conjoint Simulations, *Sawtooth Software Conference Proceedings*.

Orme, Bryan and Joel Huber (2000), Improving the Value of Conjoint Simulations, *Marketing Research*, winter 2000.