Sawtooth Software

RESEARCH PAPER SERIES

Comment on Huber: Practical Suggestions for CBC Studies

Jon Pinnell,
MarketVision Research

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

Comment on Huber:
Practical Suggestions for CBC Studies

Jon Pinnell
MarketVision Research

Joel Huber has provided a unique and very insightful recounting of the evolution of conjoint and discrete choice. I'd like to think that Joel provided the answers to the "what," "where," "why," and "when" questions of conjoint and discrete choice.

What I would like to offer is the "how" of conjoint or discrete choice, or a guide to best practices in commercial research. Inevitably, I am not able to do the "how" without including the "who." So, for the "who" — these are my perspectives based on the experimentation that we have conducted at MarketVision; and the perspectives that others have shared with me or have shared at conferences, especially at the conferences that Sawtooth has sponsored over the years. Our approaches may not be the most commonly used, and may be debated heartily, but we have adopted these as well founded as well as field tested and proved.

Most of the suggestions that follow are based on the guiding principle of increasing efficiency. The first two suggestions appear to run counter to that objective. First, let me provide some context on how we think about efficiency. There is clearly a statistical efficiency, often evaluated via D-efficiency, which expresses the statistical "goodness" of the design. However, there are other efficiencies that should be considered as well, such as the time it takes the respondent to provide a response, the amount of information in that response, and the design and analysis time required by the researcher.

Below, I offer ten practical suggestions based on our beliefs about best practices in conjoint research. They are…

Practical Suggestion #1: ***Use choice based tasks.***

The most recent *Sawtooth Solutions* (Summer 2004) published results showing what percent of studies were conducted using CBC (61%), ACA (27%), and CVA (12%). This shows that among this audience choice based designs have become the most prominent approach to conjoint analysis.

A decade ago there was a great deal of concern about the limited amount of information contained in a single choice—we knew which alternative a buyer preferred—but not by how much. We also knew that various methods (ratings versus choice) provided different answers (Huber, et al, 1992; Pinnell, 1994; Huber and Pinnell, 1995; Huber, 1997).

At the time, the tools commonly available to researchers prohibited individual-level utility estimation with choices. Instead, multinomial logit models were used to estimate utilities in aggregate, or maybe at the subgroup level. Several researchers who wanted individual level utilities suggested dual conjoint (Huisman, 1992) —with some converting ACA's individual level utilities to match those consistent with those from choice exercises (Huber, et al, 1992; Pinnell, 1994; Williams and Kilroy, 2000).

Also troubling was the difference between choices and ratings in the respondent task. Choices were a common customer and respondent task, while ratings were less intuitive and less natural.

Over the past decade, we have seen choice match in-market behavior more realistically than ratings. We have also seen the introduction of methods to allow estimation of individual level utilities from choice studies.

Choices are still less efficient than ratings, but are processed easily and quickly by respondents and appear to provide the best predictor of in-market behavior—most likely because it is the same task consumers must complete in the purchase process.

Practical Suggestion #2: ***Use randomized designs.***

Many early applications of discrete choice modeling relied on a fixed experimental design. However, with the growing popularity of computer aided interviewing over the past decade, randomized designs became feasible for a larger number of practitioners.

Mulhern (1999) found that randomized designs are nearly as efficient as fixed designs for symmetric choice experiments; for asymmetric choice experiments, randomized designs appear to be more efficient than fixed designs. In the specific example cited by Mulhern, the randomized design is 95% as efficient as the optimal fixed design for a symmetric choice experiment including ten attributes with four levels each. For an asymmetric choice experiment with eight attributes, one with seven levels, four with six levels, and three with five levels, Mulhern found the randomized design to be approximately 14% more efficient than the fixed design.

Chrzan and Orme (2000) compared the statistical efficiency of various fixed and random design strategies under different conditions. They found that one or more of the random design strategies was optimal or nearly optimal for all conditions except for a situation with "many" interactions.

The Sawtooth Software CBC Users Manual reports relative efficiencies greater than 90% for the randomized design. The median efficiency for the randomized designs reported is about 97% relative to a hypothetical orthogonal design.

Randomized designs can also reduce order and context effects relative to fixed designs. Randomized designs are less efficient than fixed orthogonal designs, but they give the researcher more flexibility and are easier to implement than fixed orthogonal designs. The loss in efficiency of randomized designs relative to fixed orthogonal designs is minimal in most choice experiments, which leads us to recommend randomized designs over fixed designs due to the gain in flexibility and ease of implementation.

There is one caveat to the recommendation of using randomized designs. It is often beneficial to include at least one fixed task as a holdout task to allow an assessment of respondent reliability.

Practical Suggestion #3: ***Use a large(r) number of alternatives per task.***

Designing choice tasks also leads to the debate over respondent fatigue and burden versus the researcher's desire to gather as much information as possible from the respondents. This leads to the questions of how many choice tasks each respondent can reliably evaluate and how many alternatives can be included in each task without overburdening the respondent.

The number of choice tasks to ask was addressed by Johnson and Orme (1996). They conclude "you can usually ask at least 20 choice tasks without degradation in data quality."

We think the number of alternatives per task is a more interesting discussion. Increasing the number of alternatives per task provides incremental information, though it may not be immediately clear how. While choices are relatively inefficient, multinomial logit is quite powerful, and it derives that power based on the number of pairwise inequalities it evaluates in a choice task. In the case of a choice task with six alternatives, five pairwise inequalities are created compared to just one pairwise inequality from the two-alternative task, or pairs. This implies statistical efficiency is greater for choice studies that include a greater number of alternatives per task. Louviere and Woodworth (1983) and Bunch, Louviere, and Anderson (1994) each conclude that paired comparison choice tasks are less efficient from a design perspective than designs that include more alternatives per task.

We evaluated (Pinnell and Englert, 1997) increasing the number of alternatives in choice tasks on cost, congruence, and efficiency criteria to determine if the added complexity of more alternatives is outweighed by higher statistical efficiency. The tasks with four alternatives took about 33% more time and those including seven alternatives took about 60% more time for respondents to evaluate than pairs. The cost in time of evaluating

choice tasks with more alternatives is relatively small compared to the gain in information collected, as summarized in the following table (based on time equalized empirical findings):

Efficiency:	Sevens nearly 6X D-efficiency of pairs
Parameters:	Sevens 59% larger than pairs
Std. Error of Parameters:	Sevens 25% smaller than pairs
Validity:	Sevens higher hit rates and lower MAE
Cost (Time):	12 sevens = 20 pairs

A secondary benefit of using a larger number of alternatives per task with randomly generated designs is that the occurrence of dominating concepts will be greatly reduced. Eliminating dominated concepts, thereby increasing the utility balance of the alternatives in the task, will increase the information contained in each respondent choice. Utility balance has been shown to increase the effectiveness of each respondent choice (Huber, Zwerina, and Pinnell, 1995; Huber and Zwerina, 1996; Johnson, Huber, and Bacon, 2003; and Johnson, Huber, and Orme, 2004).

Practical Suggestion #4: ***Only use first choices to estimate utilities.***

We have seen in previous work (Pinnell 1999a, 1999b) that respondents take longer to provide a first choice when they know that they are being asked to provide a full rank order of the alternatives compared to when they are providing only first choices. Subsequent choices, such as ranked second or ranked third did not appear useful, as those utilities were smaller (lower scale and larger error) and, more importantly, were different. There appeared to be a processing difference between first choices and later choices, with the utilities estimated from later choices showing a kink when plotted against the utilities estimated from first choice. We attributed the difference to loss aversion on the part of the respondents. The first choices from a full ranking, however, provided greater predictive validity than when respondents were only asked to provide a first choice.

For a while, our recommendation was to capture at least a second choice for each task, but to disregard all choices except for the first choice for each task. There is a cost to such questioning in terms of the additional interview length for respondents. Subsequent investigation, now using HB, has shown the superiority of first choice from a ranking question over simple first choice to be greatly diminished than with an aggregate logit.

Rank methods or allocation are still seen by some as appropriate to account for choices that are context dependent. For example, consider three possible scenarios:

- A corporate IT department might have different PC requirements for engineers than for administrative staff.

- An individual might have different preferences for beer when drinking alone versus dining out.

- A physician might have different prescribing preferences based on the patient's tolerance or susceptibility to side effects.

If there are differences in preferences, we advocate asking situation specific choice tasks. For example, ask "when purchasing computers for engineers, which would you choose?" and in a separate series of choice tasks, ask "when purchasing computers for administrative staff, which would you choose?" Estimate separate utilities for each occasion. Then, by determining the relative frequency of each occasion, the unit of analysis can be changed from a person to an occasion. We have had good success with this approach.

Practical Suggestion #5: ***Estimate utilities with HB.***

As discussed above, relying on respondent's choices rather than ratings had involved giving up individual level data. Sawtooth Software provided several methods to develop disaggregate solutions from choice data: k-logit, latent class, and ICE, but their introduction of readily available software to estimate Bayesian models truly revolutionized choice modeling.

Early applications of Bayesian methods to conjoint include Lenk, DeSarbo, Green and Young (1996), and Allenby, Arora, and Ginter (1995). The recent *Sawtooth Solutions* (Summer 2004) reports that 62% of CBC users are using HB to estimate their final model. Their adoption of HB is likely attributable to multiple success stories reported at this conference in the past.

At this conference in 2000, I reported the results (Pinnell, 2000) from six existing choice studies (summarized below) comparing hit rates using utilities developed with HB to hit rates using utilities developed with aggregate logit. In five of the six studies the results were positive, and strongly so. The anomalous sixth study was unique in that it was the only one of the six that was not a full profile study. The sixth study relied on choices from partial profile alternatives.

Comparison of Hit Rates with Disaggregation
Summary of Six Commercial Studies

	Aggregate Logit	Hierarchical Bayes	Improvement
Study One	75.8	99.5	23.8
Study Two	24.8	79.5	54.7
Study Three	60.5	62.6	2.1
Study Four	61.2	79.3	18.1
Study Five	59.2	78.8	19.6
Study Six	71.9	68.1	-3.8

While the hit rates shown above measure reliability, error in share predictions are a better measure of validity. The studies included in this analysis did not consistently include holdout tasks to allow the estimation of errors in prediction. However, others have shown improvements to share predictions from using HB.

Follow-up research (Pinnell and Fridley, 2001), which focused exclusively on partial profile designs, showed mixed results. Of the nine studies, HB showed a significant improvement in hit rate in three, no effect in two, and a statistically significant degradation in four. Our conclusion at the time was that HB was overfitting with the relatively sparse data that partial profile tasks can produce.

The data sets used in this meta-analysis were not designed specifically for this particular evaluation either and did not include hold-out tasks by which we could evaluate reductions in errors of share predictions.

However, we did find it interesting that the choice tasks with more alternatives per task did better relative to those that had fewer alternatives per task, reinforcing suggestion #3 (above).

Comparison of Hit Rates with Disaggregation
Summary of Nine Partial Profile Studies

Agg. Logit	Hier. Bayes	Diff	Std. Err	t-ratio	# of Alt./Task
73.4%	66.5%	-6.9%	0.009	-7.24	3
68.0%	65.7%	-2.3%	0.004	-5.67	3
59.2%	57.0%	-2.1%	0.006	-3.58	3
64.1%	59.6%	-4.5%	0.015	-3.05	3
56.0%	56.5%	0.6%	0.011	0.54	4
52.2%	53.6%	1.4%	0.017	0.82	4
47.6%	51.7%	4.0%	0.017	2.38	4
46.6%	57.6%	11.0%	0.015	7.41	4
32.8%	48.6%	15.8%	0.011	14.84	5

Sawtooth Software, partly in reaction to this finding and working with Peter Lenk, modified their HB software to allow the influence of the prior covariance matrix to be tuned by the researcher. After such tuning, Orme showed that the apparent detrimental effects of HB could be eliminated. However, it was not the case that each partial profile study could be improved with HB but at least parity results were achieved in each case. Which leads to the next suggestion…

Practical Suggestion #6: ***Don't use partial profile designs blindly.***

The text supporting the previous suggestion provides a cautionary tale regarding partial profile tasks, especially with disaggregate (HB) analysis.

Partial profile tasks provide a mechanism to execute choice tasks with large numbers of attributes. However, they have been shown to produce mixed results. Two new works reported in this volume (Proceedings of the 2004 Sawtooth Software Conference) show partial profile tasks to under-perform relative to full profile choice tasks.

Specifically, Frazier and Jones show that partial profile tasks produced MAE in share predictions nearly 50% larger than the standard full profile with a none option (when averaged across the three reported studies).

Separately, Johnson, Huber and Orme show hit rates for partial profile tasks to be 8 points lower than full profile tasks (71% vs. 63%), and MAEs in share predictions were more than 50% larger for partial profile than for full profile (7.1 vs. 4.6). The authors also showed that derived attribute importances from partial profile are flatter (more nearly equal) than those derived from full profile. This mirrors the result others have seen comparing utilities from ACA (which is also partial profile) to other full profile methods.

One other point on Johnson, Huber and Orme, the authors (almost apologetically) mentioned that the results predicting holdout tasks using utilities derived from partial profile tasks might have been hampered by a methodological bias as the holdouts were full profile. I personally don't believe this represents a methodological bias until we can buy partial profile products in the marketplace. One could argue that full profile stimuli produce a simplification heuristic in respondents and that might be detrimental to our ability to predict in-market behavior. However, it is not clear that the same simplification is not taking place in market. This is a topic where more research would be beneficial.

Practical Suggestion #7: ***Assume price dependence, but not linearity.***

Depending upon the category being studied, the base price of a product can vary markedly by market and channel. Moreover, within a category very similar products can have very different prices. These variations should be included in the choice design, with alternative specific pricing for each product, as well as by market and channel, as appropriate for the product category.

We are often asked by clients to include a large number of levels for the price attribute. Our inclination was often to code price as a linear variable (or to transform price such as taking the log of price). Our rationale was that by increasing the number of levels of price, ceteris paribus, we were decreasing the certainty in the parameter estimate of any one level. By solving for a single price attribute, rather than multiple parameters—one for each level—we believed we were smoothing out the error associated with any one level's estimated utility.

However, our practice has largely changed to include price as a part worth function rather than a linearly coded attribute. We have found this to replicate well and better match in-market pricing changes as well as support the notion of psychological pricing. It reinforces the finding Marder (1997) reported.

In addition, we have often found it beneficial to include exogenous variables into the price utility—specifically, price cut-off constraints. We, and others (for example, Swait, 1998; Chrzan and Boeger, 1999; Frazier and Patterson, 2000), have incorporated a cut-off penalty into the price variable when estimating utilities. As Johnson and Orme showed, respondents can change their price orientation during a choice exercise. And as Huber et al (1992) suggest, the orthogonal representation of price and brand in conjoint exercises diminishes each one's usefulness to respondents about their respective cues. Including a soft penalty creates a kink in the demand curve, which we believe better mirrors consumer behavior in reality (though not necessarily in hold-out tasks).

Practical Suggestion #8: ***Test each respondent's reliability.***

An earlier suggestion mentioned including one fixed task to test each respondent's reliability. In practice, we commonly hardcode the first task in a choice exercise and exclude it from utility estimation. This task can provide a check on the scaling of the utilities in a probabilistic choice model. This task also provides us one mechanism to test respondents' reliability. To effectively test respondent reliability, it is generally necessary to include more than one holdout task.

There is often a question about developing holdout tasks—should one make holdout tasks easy or hard for respondents? We think the answer is both. Making informed assumptions about the respondents' preferences, but without perfect information, we aim to have the most preferred alternative have about 50% greater probability of choice than the next most preferred. So for pairs, for example, we would target a 60:40 preference ratio between two alternatives.

One can gain a better understanding of respondents' reliability by repeating the same holdout and gauging respondents' reliability on the holdouts themselves. The interested reader is referred to Wittink and Johnson (1991).

To truly test respondents' reliability in practice we use re-sampling techniques. For example, we will estimate utilities using all choice tasks but the first, and use those (generally individual level) utilities to predict each respondent's choice to the first task. We will repeat this using several different tasks as the holdout. We will typically exclude 3 to 5 percent of our respondents due to poor reliability. After excluding these, we will re-run HB, though the results rarely change much.

Practical Suggestion #9: ***Beware (maybe just be aware) of sequence and order effects.***

In this volume (Proceedings of the 2004 Sawtooth Software Conference), Rogers and Renken share their results showing the importance of a more realistic shelf-like presentation. We often use a similar store shelf representation of products. We have found (not surprisingly) that the presentation of those products on the screen can influence their relative appeal. As a general rule, if we are doing a store shelf presentation of products, we will employ at least two and generally four rotations, varying what is in the upper left portion of the screen.

Also, when we are testing a wide range of prices, we will typically restrict the range of prices a respondent can see at the beginning of the choice tasks, and then broaden the range of available prices as the tasks continue. We think this is appropriate to gauge "near in" pricing and then test larger, but less likely, pricing changes.

It has been shown (Johnson and Orme 1996; Huber et al 1992) that price can change in importance during the course of a choice exercise. Therefore, building a simulator from such a sequential design involves an additional calibration step. Similarly, comparing the strength of brand preference from early tasks to late tasks will require a similar calibration.

Practical Suggestion #10: ***Further our science, and THINK!***

For all that Sawtooth Software has done to help the practicing researcher, and they surely have done much, they have yet to develop software that will think for us. Researchers, by our very nature, have a need for information and a healthy (albeit sometimes overbearing) skepticism. However, when we miss-take our purpose with the rote process, we can forget our true research objective. The fanciest of designs or most elaborate of models will not compensate for other errors or shortcomings in the design or execution of research.

These conferences that Sawtooth Software has produced have done much to develop our collective science. It continues to be up to each of us, though, to continue to test, experiment, and share findings regarding best practices with an open mind and an eager willingness to learn.

Please continue to question conventional wisdom.

References:

Allenby, Greg, Neeraj Arora, and James Ginter (1995), "Incorporating Prior Knowledge into the Analysis of Conjoint Studies," *Journal of Marketing Research*.

Bunch, David, Jordan Louviere, and Don Anderson (1994), "A Comparison of Experimental Design Strategies for Multinomial Logit Models: The Case of Generic Attributes," Working Paper, Graduate School of Business, University of California, Davis.

Chrzan, Keith, and Leesa Boeger (1999), "Improving Choice Predictions Using a Cutoff-Constrained Aggregate Choice Model," Presented at INFORMS Marketing Science Conference; Syracuse, NY.

Chrzan, Keith, and Bryan Orme (2000), "An Overview and Comparison of Design Strategies for Choice-Based Conjoint Analysis," Sawtooth Software Conference Proceedings.

Frazier, Curtis and Michael Patterson (2000), "Cutoff-Constrained Discrete Choice Models," Sawtooth Software Conference Proceedings.

Frazier, Curtis and Urszula Jones (2004), "The Effect of Design Decisions on Business Decision Making," This volume.

Huber, Joel, Dick Wittink, Richard Johnson and Richard Miller (1992), "Learning Effects in Preference Tasks: Choice Based versus Standard Conjoint," Sawtooth Software Conference Proceedings.

Huber, Joel and Jon Pinnell (1995), "Consistent Differences between Experimental Choice and Ratings-Based Tradeoffs," Presented at INFORMS Marketing Science; Sydney, NSW, Australia.

Huber, Joel, Klaus Zwerina and Jon Pinnell (1996), "Are Utility Balanced Choice Designs Really More Efficient?" Presented at INFORMS Marketing Science Conference; Gainesville, FL.

Huber, Joel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*.

Huber, Joel (1997), "What We Have Learned from 20 Years of Conjoint Research: When to Use Self-Explicated, Graded Pairs, Full Profile or Choice Experiments," Sawtooth Software Conference Proceedings.

Huisman, Dirk (1992), "Price-Sensitivity Measurement of Multi-Attribute Products," Sawtooth Software Conference Proceedings.

Johnson, Richard and Bryan Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint Studies?" Presented at AMA's ART Forum, Beaver Creek, CO.

Johnson, Richard, Joel Huber, and Lynd Bacon (2003), Adaptive Choice-Based Conjoint, Sawtooth Software Conference Proceedings.

Johnson, Richard, Joel Huber, and Bryan Orme (2004), "A Second Test of Adaptive Choice-Based Conjoint Analysis (The Surprising Robustness of Standard CBC Designs," This volume.

Lenk, Peter, Wayne DeSarbo, Paul Green, and Martin Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Part-worth Heterogeneity from Reduced Experimental Designs," *Marketing Science*.

Louviere, Jordon, and George Woodworth (1983), "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data." *Journal of Marketing Research*.

Marder, Eric (1997), *The Laws of Choice: Predicting Customer Behavior*. Free Press.

Mulhern, Michael (1999), "Assessing the Relative Efficiency of Fixed and Randomized Experimental Designs," Sawtooth Software Conference Proceedings.

Pinnell, Jon (1994), "Multistage Conjoint Methods to Measure Price Sensitivity." Presented at AMA's ART Forum, Beaver Creek, CO.

Pinnell, Jon and Sherry Englert (1997), "Number of Choice Alternatives in Discrete Choice Modeling," Sawtooth Software Conference Proceedings.

Pinnell, Jon (1999a), "Depth of Probing, Allocation and Rank Order Methods in Consumer Choice Analysis," Presented at INFORMS Marketing Science Conference; Syracuse, NY.

Pinnell, Jon (1999b), "Should Choice Researchers Always Use 'Pick One' Respondent Tasks?" Sawtooth Software Conference Proceedings.

Pinnell, Jon (2000), "Customized Choice Designs: Incorporating Prior Knowledge and Utility Balance in Choice Experiments," Sawtooth Software Conference Proceedings.

Pinnell, Jon and Lisa Fridley (2001), "The Effects of Disaggregation with Partial Profile Choice Experiments," Sawtooth Software Conference Proceedings.

Rogers, Greg and Tim Renken (2004), "The Importance of Shelf Presentation in Choice-Based Conjoint Studies," This volume.

Swait, Joffre (1998), “A Model of Heuristic Decision Making: The Role of Cutoffs in Choice Processes,” Working Paper, University of Florida.

Willaims, Peter and Denis Kilroy (2000), “Calibrating Price in ACA: The ACA Price Effect and How to Manage It,” Sawtooth Software Conference Proceedings.

Wittink, Dick and Richard Johnson (1991), “Estimating the Agreement Between Choices Among Discrete Objects and Conjoint-Ratings-Based Predictions After Correcting for Attenuation,” Working Paper, Cornell University.