PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

October 2013

RESEARCH SPACE AND REALISTIC PRICING IN SHELF LAYOUT CONJOINT (SLC)

PETER KURZ[1]
TNS INFRATEST
STEFAN BINNER
BMS MARKETING RESEARCH + STRATEGY
LEONHARD KEHL
PREMIUM CHOICE RESEARCH & CONSULTING

WHY TALK ABOUT SHELVES?

For consumers, times long ago changed. Rather than being served by a shop assistant, super- and hypermarkets have changed the way we buy products, especially fast-moving consumer goods (FMCGs). In most developed countries there is an overwhelming number of products for consumers to select from:

"Traditional Trade"

"Modern Trade"

As marketers became aware of the importance of packaging design, assortment and positioning of their products on these huge shelves, researchers developed methods to test these new marketing mix elements. One example is a "shelf test" where respondents are interviewed in front of a real shelf about their reaction to the offered products. (In FMCG work, the products are often referred to as "stock keeping units" or "SKUs," a term that emphasizes that each variation of flavor or package size is treated as a different product.)

For a long time, conjoint analysis was not very good at mimicking such shelves in choice tasks: early versions of CBC were limited to a small number of concepts to be shown. Furthermore the philosophical approach for conjoint analysis, let's call it the traditional conjoint approach, was driven by taking products apart into attributes and levels.

However, this traditional approach missed some key elements in consumers' choice situation in front of a modern FMCG shelf, e.g.:

- How does the packaging design of an SKU communicate the benefits (attribute levels) of a product?

[1] Correspondence: Peter Kurz, Head of Research & Development TNS Infratest (Peter.Kurz@tns-global.com)
Stefan Binner, Managing Director, bms marketing research + strategy (s.binner@bms-net.de)
Leonhard Kehl, Managing Director, Premium Choice Research & Consulting (leonhard.kehl@premium-choice-research.com)

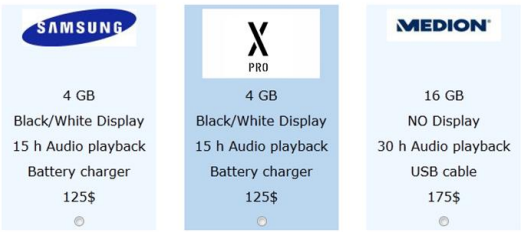
- How does an SKU perform in the complex competition with the other SKUs in the shelf?

As it became easy for researchers to create shelf-like choice tasks (in 2013, among Sawtooth Software users who use CBC, 11% of their CBC projects employed shelf display) a new conjoint approach developed: “Shelf Layout Conjoint” or “SLC.”

How is Shelf Layout Conjoint different?

The main differences between Traditional and Shelf Layout Conjoint are summarized in this chart:

TRADITIONAL CONJOINT



- Products or concepts consist usually of defined attribute levels
- More rational or textual concept description (compared to packaging picture)
- Almost no impact of package design
- Usually not too many concepts per task

➔many attributes—few concepts

SHELF LAYOUT CONJOINT

- Communication of “attributes” through non-varying package design (instead of levels)
- Visibility of all concepts at once
- Including impact of assortment
- Including impact of shelf position and number of facings
- Information overflow

➔ few visible attributes (mainly product and price—picture only)—many concepts

Many approaches are used to represent a shelf in a conjoint task. Some are very simple:

Some are quite sophisticated:

However, even the most sophisticated computerized visualization does not reflect the real situation of a consumer in a supermarket (Kurz 2008). In that paper, comparisons between a simple grid of products from which consumers make their choices and attempts to make the choice exercise more realistic by showing a store shelf in 3D showed no significant differences in the resulting preference share models.

The Challenges of Shelf Layout Conjoint

Besides differences in the visualization of the shelves, there are different objectives SLCs can address, including:

- pricing
- product optimization
- portfolio optimization
- positioning
- layout
- promotion

SLCs also differ in the complexity of their models and experimental designs, ranging from simple main effects models up to complex Discrete Choice Models (DCM's) with lots of attributes and parameters to be estimated.

Researchers often run into very complex models, with one attribute with a large number of levels (the SKU's) and related to each of these levels one attribute (often, price) with a certain number of levels. Such designs could easily end up with several hundred parameters to be estimated. Furthermore, for complex experimental designs, layouts have to be generated in a special way, in order to retain realistic relationships between SKUs and realistic results. So-called "alternative-specific designs" are often used in SLC, but that does not necessarily mean that it is always a good idea to estimate price effects as being alternative-specific. In terms of estimating utility values (under the assumption you estimate interaction effects, which lead to alternative-specific price effects), many different coding-schemes can be prepared which are mathematically identical. But, the experimental design behind the shelves is slightly different. Different design strategies affect how much level overlap occurs and therefore how efficient the

estimation of interactions can be. Good strategies to reduce this complexity in the estimation stage are crucial.

With Shelf Layout Conjoint now easily available to every CBC user, we would like to encourage researchers to use this powerful tool. However, there are at least five critical questions in the design of Shelf Layout Conjoints which need to be addressed:

- Are the research objectives suitable for Shelf Layout Conjoint?
- What is the correct target group and SKU space (the "research space")?
- Are the planned shelf layout choice tasks meaningful for respondents, and will they provide the desired information from their choices?
- Can we assure realistic inputs and results with regard to pricing?
- How can we build simulation models that provide reliable and meaningful results?

As this list suggests, there are many topics, problems and possible solutions with Shelf Layout Conjoint. However, this paper focuses on only three of these very important issues. We take a practitioner's, rather than an academic's point of view. The three key areas we will address are:

1. Which are suitable research objectives?
2. How to define the research space?
3. How to handle pricing in a realistic way?

Suitable Research Objectives for Shelf Layout Conjoint

Evaluating suitable objectives requires that researchers be aware of all the limitations and obstacles Shelf Layout Conjoint has. So, we begin by introducing three of those key limitations.

1. Visualization of the test shelf. Real shelves always look different than test shelves.

Furthermore there is naturally a difference between a 21″ Screen and a 10 meter shelf in a real supermarket. The SKUs are shown much smaller than in reality. One cannot touch and feel them. 3D models and other approaches might help, but the basic issue still remains.

2. Realistic choices for consumers. Shelf Layout Conjoint creates an artificial distribution and awareness: All products are on the shelf; respondents are usually asked to look at or consider all of them.

In addition, we usually simplify the market with our test shelf. In reality every distribution chain has different shelf offerings, which might further vary with the size of the individual store. In the real world, consumers often leave store A and go to store B if they do not like the offering (shop hopping). Sometimes products are out of stock, forcing consumers to buy something different.

3. Market predictions from Shelf Layout Conjoint. Shelf Layout Conjoint provides results from a single purchase simulation. We gain no insights about repurchase (did they like the product at all?) or future purchase frequency.

In reality, promotions play a big role, not only in the shelf display but in other ways, for example, with second facings. It is very challenging to measure volumetric promotion effects such as "stocking up" purchases, but those play a big role in some product categories (Eagle 2010; Pandey, Wagner 2012).

The complexity of "razor and blade" products, where manufacturers make their profit on the refill or consumable rather than on the basic product or tool, are another example of difficult obstacles researchers can be faced with.

Suitable Objectives

Despite these limitations and obstacles Shelf Layout Conjoint can provide powerful knowledge. It is just a matter of using it to address the right objectives; if you use it appropriately, it works very well!

Usually suitable objectives for Shelf Layout Conjoint fall in the areas of either optimization of assortment or pricing.

The optimization of assortment most often refers to such issues as:

Line extensions with additional SKUs

- What is the impact (share of choice) of the new product?
- Where does this share of choice come from (customer migration and/or cannibalization)?
- Which possible line extension has the highest consumer preference or leads to the best overall result for the total brand assortment or product line?

Re-launch or substitution of existing SKUs

- What is the impact (share of choice) for the re-launch?
- Which possible re-launch alternative has the highest consumer preference?
- Which SKU should be substituted for?
- How does the result of the re-launch compare to a line extension?

Branding

- What is the effect of re-branding a line of products?
- What is the effect of the market entry of new brands or competitors?

The optimization of pricing most often involves questions like:

Price positioning

- What is the impact of different prices on share of choice and profit?
- How should the different SKUs within an assortment be priced?
- How will the market react to a competitor's price changes?

Promotions

- What is the impact (sensitivity) to promotions?
- Which SKUs have the highest promotion effect?
- How much price reduction is necessary to create a promotion effect?

Indirect pricing

- What is the impact of different contents (i.e., package sizes) on share of choice and profit?
- How should the contents of different SKUs within an assortment be defined?
- How will the market react to competitors' changes in contents?

On the other hand there are research objectives which are problematic, or at least challenging, for Shelf Layout Conjoint. Some examples include:

- Market size forecasts for a sales period
- Volumetric promotion effects
- Multi category purchase/TURF-like goals
- Positioning of products on the shelf
- Development of package design
- Evaluation of new product concepts

- New product development

Not all of the above research objectives are impossible, but they at least require very tailored or cautious approaches.

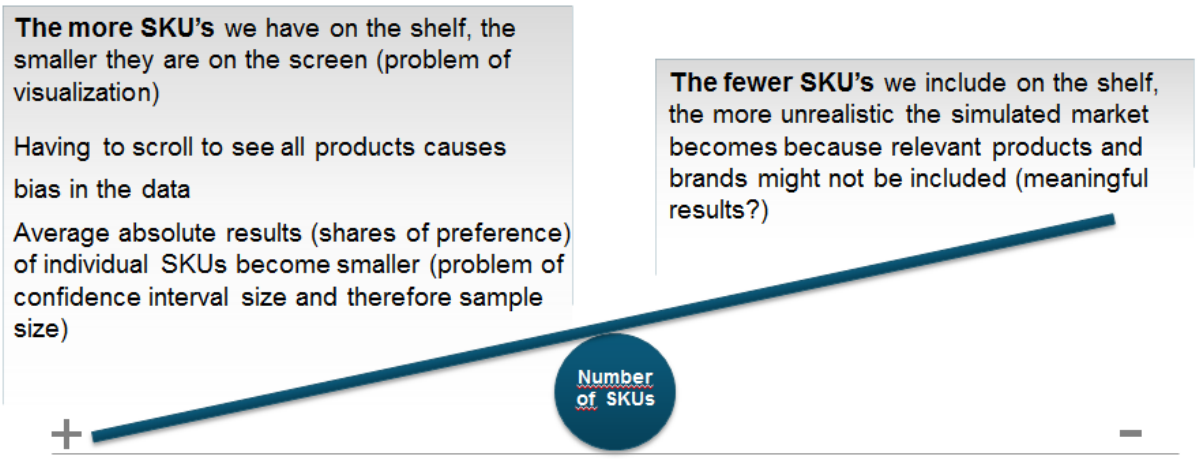
Definition of the Correct Market Scope

By the terminology "market scope" we mean the research space of the Shelf Layout Conjoint. Market scope can be defined by three questions, which are somewhat related to each other:

What SKUs do we show on the Shelf?	=> **SKU Space**
What consumers do we interview?	=> **Target Group**
What do we actually ask them to do?	=> **Context**

SKU Space

The basic problem for the definition of the SKU space:



Possible solutions to this basic problem depend heavily on the specific market and product category. Two main types of solutions are:

1. Focusing the SKU space on or by *market* segments

Such focus could be achieved by narrowing the SKU space to just a part of the market such as
- distribution channel (no shop has all SKUs)
- product subcategories (e.g., product types such as solid vs. liquid)
- market segments (e.g., premium or value for money)

This will provide more meaningful results for the targeted market segment. However it may miss migration effects from (or to) other segments. Furthermore such segments might be quite artificial from the consumer's point of view.

Alternatively one could focus on the most relevant products only (80:20 rule).

2. Strategies to cope with too many SKUs

When there are more SKUs than can be shown on the screen or understood by the respondents, strategies might include

- Prior SKU selection (consideration sets)
- Multiple models for subcategories
- Partial shelves

Further considerations for the SKU space:

- Private labels (which SKUs could represent this segment?)
- Out of stock effects (whether and how to include them?)

Target Group

The definition of the target group must align with the SKU space. If there is a market segment focus, then obviously the target group should only include customers in that segment. Conversely, if there are strategies for huge numbers of SKUs all relevant customers should be included.

There are still other questions about the target group which should also be addressed, including:

- Current buyers only or also non-buyers?
- Quotas on recently used brands or SKUs?
- Quotas on distribution channel?
- Quotas on the purchase occasion?

Context

Once the SKU space and the target group are defined the final element of "market scope" is to create realistic, meaningful choice tasks:

1. Setting the scene:
   - Introduction of new SKUs
   - Advertising simulation
2. Visualization of the shelf
   - Shelf Layout (brand blocks, multiple facings)
   - Line pricing/promotions
   - Possibility to enlarge or "examine" products
3. The conjoint/choice question
   - What exactly is the task?
   - Setting a purchase scenario or not?
   - Single choice or volumetric measurement?

PRICING

Pricing is one of the most, if not the most, important topic in Shelf Layout Conjoint. In nearly all SLCs some kind of pricing issue is included as an objective. But "pricing" does not mean just one common approach. Research questions in regard to pricing are very different between different studies. They start with easy questions about the "right price" of a single SKU. They often include the pricing of whole product portfolios, including different pack sizes, flavors and

variants and may extend to complicated objectives like determining the best promotion price and the impact of different price tags.

Before designing a Shelf Layout Conjoint researchers must therefore have a clear answer to the question: “How can we obtain realistic input on pricing?”

Realistic pricing does not simply mean that one needs to design around the correct regular sales price. It also requires a clear understanding of whether the following topics play a role in the research context.

Topic 1: Market Relevant Pricing

The main issue of this topic is to investigate the context in which the pricing scenario takes place. Usually such an investigation starts with the determination of actual sales prices. At first glance, this seems very easy and not worth a lot of time. However, most products are not priced with a single regular sales price. For example, there are different prices in different sales channels or store brands. Most products have many different actual sales prices. Therefore one must start with a closer look at scanner data or list prices of the products in the SKU space.

As a next step, one has to get a clear understanding of the environment in which the product is sold. Are there different channels like hypermarkets, supermarkets, traders, etc. that have to be taken into account? In the real world, prices are often too different across channels to be used in only one Shelf Layout Conjoint design. So we often end up with different conjoint models for the different channels.

Furthermore, the different store brands may play a role. Store brand A might have a different pricing strategy because it competes with a different set of SKUs than store brand B. How relevant are the different private labels or white-label/generic products in the researched market?

In consequence one often ends up with more than one Shelf Layout Conjoint model (perhaps even dozens of models) for one “simple” pricing context. In such a situation, researchers have to decide whether to simulate each model independently or to build up a more complex simulator. This will allow pricing simulations on an overall market level, tying together the large number of choice models, to construct a realistic “playground” for market simulations.

Topic 2: Initial Price Position of New SKUs

With the simulated launch of new products one has to make prior assumptions about their pricing before the survey is fielded. Thus, one of the important tasks for the researcher and her client is to define reasonable price points for the new products in the model. The price range must be as wide as necessary, but as narrow as possible.

Topic 3: Definition of Price Range Widths and Steps

Shelf Layout Conjoint should cover all possible pricing scenarios that would be interesting for the client. However, respondents should not be confronted with unrealistically high or low prices. Such extremes might artificially influence the choices of the respondent and might have an influence on the measured price elasticity. Unrealistically high price elasticity is usually caused by too wide a price range, with extremely cheap or extremely expensive prices. One should be aware that the price range over which an SKU is studied has a direct impact on its elasticity results! This is not only true for new products, where respondents have no real price

knowledge, but also for existing products. Furthermore unrealistically low or high price points can result in less attention from respondents and more fatigue in the answers of respondents, than realistic price changes would have caused.

Topic 4: Assortment Pricing (Line Pricing)

Many clients have not just a single product in the market, but a complete line of competing products on the same shelf. In such cases it is often important to price the products in relation to each other.

A specific issue in this regard is line pricing: several products of one supplier share the same price, but differ in their contents (package sizes) or other characteristics. Many researchers measure the utility of prices independently for each SKU and create line pricing only in the simulation stage. However, in this situation, it is essential to use line-priced choice tasks in the interview: respondents' preference structure can be very different when seeing the same prices for all products of one manufacturer rather than seeing different prices, which often results in choosing the least expensive product. This leads to overestimation of preference shares for cheaper products.

A similar effect can be observed if the relative price separations of products are not respected in the choice tasks. For example: if one always sells orange juice for 50 cents more than water, this relative price distance is known or learned by consumers and taken into account when they state their preference in the choice tasks.

Special pricing designs such as line pricing can be constructed by exporting the standard design created with Sawtooth Software's CBC into a CSV format and reworking it in Excel. However, one must test the manipulated designs afterwards in order to ensure the prior design criteria are still met. This is done by re-importing the modified design into Sawtooth Software's CBC and running efficiency tests.

Topic 5: Indirect Pricing

In markets where most brands offer line pricing the real individual price positioning of SKUs is often achieved through variation in their package content sizes. This variation can be varied and modeled in the same way and with the same limitations as monetary prices. However, one must ensure that the content information is sufficiently visible to the consumers (e.g., written on price tags or on the product images).

Topic 6: Price Tags

Traditionally, prices in conjoint are treated like an attribute level and are simply displayed beneath each concept. Therefore in many Shelf Layout Conjoint projects the price tag is simply the representation of the actual market price and the selected range around it. However, in reality consumers see the product name, content size, number of applications, price per application or

standard unit in addition to the purchase price. (In the European Community, such information is mandatory by law; in many other places, it is at least customary if not required.)

In choice tasks, many respondents therefore search not only for the purchase price, but also for additional information about the SKUs in their relevant choice set. Oversimplification of price tags in Shelf Layout Conjoint does not sufficiently reflect the real decision process. Therefore, it is essential to include the usual additional information to ensure realistic choice tasks for respondents.

Topic 7: Promotions

The subject of promotions in Shelf Layout Conjoint is often discussed but controversial. In our opinion, only some effects of promotions can be measured and modeled in Shelf Layout Conjoint. SLC provides a one-point-in-time measurement of consumer preference. Thus, promotion effects which require information over a time period of consumer choices cannot be measured with SLC. It is essential to keep in mind that we can neither answer the question if a promotion campaign results in higher sales volume for the client nor make assumptions about market expansion—we simply do not know anything about the purchase behavior (what and how much) in the future period.

However, SLC can simulate customers' reaction to different promotion activities. This includes the simulation of the necessary price discount in order to achieve a promotion effect, comparison of the effectiveness of different promotion types (e.g., buy two, get one free) as well as simulation of competitive reactions, but only at a single point in time. In order to analyze such promotion effects with high accuracy, we recommend applying different attributes and levels for the promotional offers from those for the usual market prices. SLC including promotion effects therefore often have two sets of price parameters, one for the regular market price and one for the promotional price.

Topic 8: Price Elasticity

Price Elasticity is a coefficient which tells us how sales volume changes when prices are changed. However, one cannot predict sales figures from SLC. What we get is "share of preference" or "share of choice" and we know whether more or fewer people are probably purchasing when prices change.

In categories with single-unit purchase cycles, this is not much of a problem, but in the case of Fast Moving Consumer Goods (FMCG) with large shelves where consumers often buy more than one unit—especially under promotion—it is very critical to be precise when talking about price elasticity. We recommend speaking carefully of a "price to share of preference coefficient" unless sales figures are used in addition to derive true price elasticity.

The number of SKUs included in the research has a strong impact on the "price to share of preference coefficient." The fewer SKUs one includes in the model, the higher the ratio; many researchers experience high "ratios" that are only due to the research design. But they are certainly wrong if the client wants to know the true "coefficient of elasticity" based on sales figures.

Topic 9: Complexity of the Model

SLC models are normally far more complex than the usual CBC/DCM models. The basic structure of SLC is usually one many-leveled SKU attribute and for each (or most) of its levels, one price attribute. Sometimes there are additional attributes for each SKU such as promotion or content. As a consequence there are often too many parameters to be estimated in HB. Statistically, we have "over-parameterization" of the model. However there are approaches to reduce the number of estimated parameters, e.g.:

- Do we need part-worth estimates for each price point?
- Could we use a linear price function?
- Do we really need price variation for all SKUs?
- Could we use fixed price points for some competitors' SKUs?
- Could we model different price effects by price tiers (such as low, mid, high) instead of one price attribute per SKU?

Depending on the quantity of information one can obtain from a single respondent, it may be better to use aggregate models than HB models. The question is, how many tasks could one really ask of a single respondent before reaching her individual choice task threshold, and how many concepts could be displayed on one screen (Kurz, Binner 2012)? If it's not possible to show respondents a large enough number of choice tasks to get good individual information, relative to the large number of parameters in an SLC model, HB utility estimates will fail to capture much heterogeneity anyway.

Topics beyond this Paper

How can researchers further ensure that Shelf Layout Conjoint provides reliable and meaningful results? Here are some additional topics:

- Sample size and number of tasks
- Block designs
- Static SKUs
- Maximum number of SKUs on shelf
- Choice task thresholds
- Bridged models
- Usage of different (more informative) priors in HB to obtain better estimates

Eight Key Take-Aways for SLC

1. Be aware of its limitations when considering Shelf Layout Conjoint as a methodology for your customers' objectives. One cannot address every research question.

2. Try hard to ensure that your pricing accurately reflects the market reality. If one model is not possible, use multi-model simulations or single market segments.
3. Be aware that the price range definition for a SKU has a direct impact on its elasticity results and define realistic price ranges with care.
4. Adapt your design to the market reality (e.g., line pricing), starting with the choice tasks and not only in your simulations.
5. Do not oversimplify price tags in Shelf Layout Conjoint; be sure to sufficiently reflect the real decision environment.
6. SLC provides just a one point measurement of consumer preference. Promotion effects that require information about a time period of consumer choices cannot be measured.
7. Price elasticity derived from SLC is better called "price to share of preference coefficients."
8. SLC often suffers from "over-parameterization" within the model. One should evaluate different approaches to reduce the number of estimated parameters.

Peter Kurz

Stefan Binner

Leonhard Kehl

References

Eagle, Tom (2010): Modeling Demand Using Simple Methods: Joint Discrete/Continuous Modeling; *2010 Sawtooth Software Conference Proceedings.*

Kehl, Leonhard; Foscht, Thomas; Schloffer, Judith (2010): Conjoint Design and the impact on Price-Elasticity and Validity; *Sawtooth Europe Conference Cologne.*

Kurz, Peter; Binner, Stefan (2012): "The Individual Choice Task Threshold" Need for Variable Number of Choice Tasks; *2012 Sawtooth Software Conference Proceedings.*

Kurz, Peter (2008): A comparison between Discrete Choice Models based on virtual shelves and flat shelf layouts; *SKIM Working Towards Symbioses Conference Barcelona.*

Orme, Bryan (2003): Special Features of CBC Software for Packaged Goods and Beverage Research; *Sawtooth Software*, via Website.

Pandey, Rohit; Wagner, John; Knappenberger, Robyn; (2012): Building Expandable Consumption into a Share-Only MNL Model; *2012 Sawtooth Software Conference Proceedings.*

confirm. This would create a different mental frame making people think about other, possibly bad or less important, experiences to answer that question.

The subsets of memories from the two questions might be different or overlap partially. We believe that there is an overlap and, thus, the differences in preference parameters between the two questions can be represented by the change in scale. This is a scale parameter λ (question framing effect) in our model. However, if the retrievals in the two questions are independent and generate different samples of memories, then a model where we allow for independent preference β parameters would perform better than the model that only adjusts the scale parameter.

The third component of the model is related to the error term distribution in models for Best-Worst choices tasks. Traditionally, a logit specification that is based on the maximum extreme value assumption of the error term is used. This is mostly due to the mathematical and computational convenience of these models: the probability expressions have closed forms and, hence, the model can be estimated relatively fast.

We, however, want to give the error term distributional assumption serious consideration by thinking about more appropriate reasons for the use of extreme value (asymmetric) versus normal (symmetric) distributional assumptions. The question is: can we use the psychology literature to help us justify the use of one specification versus another? As an example of how it can be done, we use the theory of episodic versus semantic memory retrieval and processing (Tulvin, 1972).

When answering Best-Worst questions, people need to summarize the subsets of information that were just retrieved from memory. If the memories and associations are aggregated by averaging (or summing) over the episodes and experiences (which would be consistent with a semantic information processing and retrieval mechanism), then that would be consistent with the use of the normally distributed (symmetric) error term due to the Central Limit Theorem. However, if respondents pay attention to specific episodes within these samples of information looking for the most representative episodes to answer the question at hand (which would be consistent with an episodic memory processing mechanism), then the extreme value error term assumption would be justified. This is due to Extreme Value Theory, which says that the maximum, or minimum, of a random variable is distributed Max, or Min, extreme value. Thus, in the “select-the-best” decision it is appropriate to use the maximum extreme value error term, and in the “select-the-worst” question, the minimum extreme value distribution is justified.

Equation 1 is the model for one Best-Worst decision task. This equation shows the model based on the episodic memory processing, or extreme value error terms. It includes two possible sequences, that are indexed by parameter θ, order scale parameter ψ in the second decision, exclusion of the first choice from the set in the second decision, and our question framing scaling parameter λ. The model with the normal error term assumption has the same conceptual structure but the choice probabilities have different expressions.

Equation 1. Sequential Evaluation Model (logit specification)

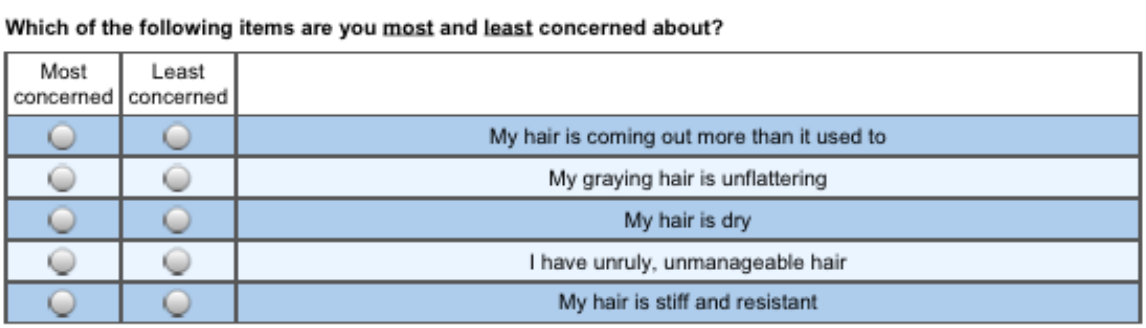
$$\Pr\left(y_{best}=i, y_{worst}=j|\beta,\lambda,\psi,\theta\right)=$$
$$=\theta\frac{\exp\left(x_i\beta\right)}{\sum\limits_k \exp\left(x_k\beta\right)}\frac{\exp\left(-x_j\psi\lambda\beta\right)}{\sum\limits_{l,-i}\exp\left(-x_l\psi\lambda\beta\right)}+\left(1-\theta\right)\frac{\exp\left(-x_j\lambda\beta\right)}{\sum\limits_l \exp\left(-x_l\lambda\beta\right)}\frac{\exp\left(x_i\psi\beta\right)}{\sum\limits_{k,-j}\exp\left(x_k\psi\beta\right)}$$

This model is a generalized model and includes some existing models as special cases. For example, if we use the probability weight instead of sequence indicator θ, then under specific values of that parameter, our model would include the traditional MaxDiff model. The concordant model by Marley et al. (2005) would also be a special case of our modified model.

EMPIRICAL APPLICATION AND RESULTS

We applied our model to data that was collected from an SSI panel. Respondents went through 15 choices tasks with five items each as is shown in Figure 2. The items came from a list of 15 hair care concerns and issues. We analyzed responses from 594 female respondents over 50 years old. This sample of the population is known for high involvement with the hair care category. For example, in our sample, 65% of respondents expressed some level of involvement with the category.

Figure 2. Best-Worst task



We estimated our proposed models with and without the proposed effects. We used Hierarchical Bayesian estimation where preference parameters β, order effect ψ and context effect λ are heterogeneous. To ensure empirical identification, the latent sequence parameter θ is estimated as an indicator parameter from Bernoulli distribution and is assumed to be the same for all respondents. We use standard priors for the parameters of interest.

Table 1 shows the improvement of model fit (log marginal density, Newton-Raftery estimator) as the result of the presence of each effect, that is, the marginal effect of each model element. Table 2 shows in-sample and holdout hit probabilities for the Best-Worst pair (random chance is 0.05).

Table 1. Model Fit

	LMD NR
Exploded logit (single evaluation)	-13,040
Context effect only	-12,455
Order effect only	-11,755
Context and order effects together	-11,051

These tables show significant improvement in fit from each of the components of the model. The strongest improvement comes from the order effect, indicating that the sequential mechanisms we assumed are more plausible given the data than the model with the assumption of single evaluation. The context effect improves the fit as well, indicating that it is likely that the two questions, "select-the-best" and "select-the-worst," are processed differently by respondents. The model with both effects included into the model is the best model not just with respect to fit to the data in-sample, but also in terms of holdout performance.

Table 2. Improvement in Model Fit

	In-sample Hit Probabilities	*In-sample Improvement* *		*Holdout Hit Probabilities*	*Holdout Improvement* *
Exploded logit (single evaluation)	0.3062	-		0.2173	-
Context effect only	0.3168	3.5 %		0.2226	2.4%
Order effect only	0.3443	12.4%		0.2356	8.4%
Context and order effects together	0.3789	23.7%		0.2499	15.0%

* Improvements are calculated over the metric in first line, which comes from the model that assumes single evaluation (ranking) in Best-Worst tasks.

We found that both error term assumptions (symmetric and asymmetric) are plausible as the fit of the models are very similar. Based on that finding, we can recommend using our sequential logit model, as it has computational advantages over the sequential probit model. The remaining results we present are based on the sequential logit model.

We also found that the presence of dependent preference parameters between the "Best" and "Worst" questions (question framing scale effect λ) is a better fitting assumption than the assumption of independence of β's from the two questions.

From a managerial standpoint, we want to show why it is important to use our sequential evaluation model instead of single evaluation models. We compared individual preference parameters from two models: our best performing model and exploded logit specification (single-evaluation ranking model). Table 3 shows the proportion of respondents for whom a subset of top items is the same between these two models. For example, for the top 3 items related to the hair care concerns and issues, the two models agree only for 61% of respondents. If we take into account the order within these subsets, then the matching proportion drops to 46%. This means that for more than half of respondents in our study, the findings and recommendations will be different between the two models. Given the fact that our model of

sequential evaluation is a better fitting model, we suggest that the results from single evaluation models can be misleading for managerial implications and that the results from our sequential evaluation model should be used.

Table 3. Proportion of respondents matched on top *n* items of importance between sequential and single evaluation (exploded logit) models

Top n items	*Proportion of respondents (Order does not matter)*	*Proportion of respondents (Order does matter)*
1	83.7%	83.7%
2	72.1%	65.0%
3	61.1%	46.5%
4	53.2%	29.1%
5	47.0%	18.4%
6	37.7%	10.4%

Our sequential evaluation model also provides additional insights about the processes that are present in Best-Worst choice tasks. First, we found that in these tasks respondents are more likely to eliminate the worst alternative from the list and then select the best one. This is consistent with literature that suggests that people, when presented with multiple alternatives, are more likely to simplify the task by eliminating, or screening-out, some options (Ordonez, 1999; Beach and Potter, 1992). Given the nature of the task in our application, where respondents had to select the most important and least important items, it is not surprising that eliminating what is not important first would be the most likely strategy.

This finding, however, is in contrast with the click data that was collected in these tasks. We found that about 68% of clicks were best-then-worst. To understand this discrepancy, we added the observed sequence information into our model by substituting the indicator of latent sequence θ with the decision order that we observed. Table 4 shows the results of the fit of these models. The data on observed sequence makes the fit of the model worse. This suggests that researchers need to be careful when thinking that click data is a good representation of the latent processes driving consumer decisions in Best-Worst choice tasks.

Table 4. Fit of the models with latent and observed sequence of decisions.

	LMD NR	*In-sample Hit Probabilities*	*Holdout Hit Probabilities*
Latent sequence	-11,051	0.3789	0.2499
Observed sequence	-12,392	0.3210	0.2247

To investigate order further, we manipulated the order of the decisions by collecting responses from two groups. One group was forced to select the best alternative first and then select the worst, and the second group was forced to select in the opposite order. We found that the fit of our model with the indicator of latent sequence is the same as for the group that was required to select the worst alternative first. This analysis gives us more confidence in our

finding. To understand why the click data seem to be inconsistent with the underlying decision making processes in respondents' minds is outside of the scope of this paper, but is an important topic for future research.

Our model also gives us an opportunity to learn about other effects present in Best-Worst choice tasks and account for those effects. For instance, there is a difference in the certainty level between the first and the second decisions. As expected, the second decision is less error prone than the first. The mean of the posterior distribution of the order effect ψ is greater than one for almost all respondents. This finding is consistent with our expectation that the decrease in the difficulty of the task in the second choice will impact the certainty level. While we haven't directly tested the impact of the number of items on the list on certainty level, our finding is expected.

Another effect that we have included in our model is the scale effect of question framing λ, which represents the level of certainty in the parameters that a researcher obtains between the best and worst selections as the result of the response elicitation procedure—"best" versus "worst." We found that the average of the sample for this parameter is 1.17, which is greater than one. This means that, on average, respondents in our sample are more consistent in their "worst" choices. However, we found significant heterogeneity in this parameter among respondents.

To understand what can explain the heterogeneity in this parameter, we performed a post-estimation analysis of the context scale parameter as it relates to an individual's expertise level, which was also collected in the survey. We found a negative correlation of -0.16 between the means of the context effect parameter and the level of expertise, meaning that experts are likely to be more consistent in what is important to them and non-experts are more consistent about what is not important to them.

We also found a significant negative correlation (-0.20) between the direct measure of the difficulty of the "select-the-worst" items and the context effect parameter, indicating that if it was easier to respond to the "select-the-worst" questions, λ was larger which is consistent with our proposition and expectations.

Conclusions

In this paper, we proposed a model to analyze data from Best-Worst choice tasks. We showed how the development of model specification could be driven by theories from the psychology literature. We took a deep look at how we can think about the possible processes that underlie decisions in these tasks and how to reflect that in the mathematical representation of the data-generating mechanism.

We found that our proposed model of sequential evaluation is a better fitting model than the currently used models of single evaluation. We showed that adding the sequential nature to the model specification allows other effects to be taken into consideration. We found that the second decision is more certain than the first decision, but the "worst" decision is, on average, more certain.

Finally, we demonstrated the managerial implications of the proposed model. Our model that takes into account psychological processes within Best-Worst choice tasks gives different results about what is most important to specific respondents. This finding has direct implications for

new product development initiatives and understanding the underlying needs and concerns of customers.

Greg Allenby

REFERENCES

Bacon, L., Lenk, P., Seryakova, K., Veccia, E. (2007) "Making MaxDiff More Informative: Statistical Data Fusion by Way of Latent Variable Modeling," *Sawtooth Software Conference Proceedings*, 327–343.

Beach, L. R., & Potter, R. E. (1992) "The pre-choice screening of options," *Acta Psychologica*, 81(2), 115–126.

Finn, A. & Louviere, J. (1992). "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety," *Journal of Public Policy & Marketing*, Vol. 11, No. 2 (Fall, 1992), 12–25.

Hoch, S. J. and Ha, Y.-W. (1986) "Consumer Learning: Advertising and the Ambiguity of Product Experience," *Journal of Consumer Research* , Vol. 13, 221–233.

Marley, A. A. J. & Louviere, J.J. (2005). "Some Probabilistic Models of Best, Worst, and Best-Worst Choices," *Journal of Mathematical Psychology*, 49, 464–480.

Marley, A. A. J., Flynn, T.N. & Louviere, J.J. (2008). "Probabilistic Models of Set-dependent and attribute level Best-Worst Choice," *Journal of Mathematical Psychology*, 52, 281–296.

Ordoñez, L. D., Benson, III, L. and Beach, L. R. (1999), "Testing the Compatibility Test: How Instructions, Accountability, and Anticipated Regret Affect Prechoice Screening of Options," *Organizational Behavior and Human Decision Processes*, Vol. 78, 63–80.

Snyder, M. (1981) "Seek and ye shall find: Testing hypotheses about other people," in C. Heiman, E. Higgins and M. Zanna, eds, 'Social Cognition: The Ontario Symposium on Personality and Social Psychology,' Hillside, NJ: Erlbaum, 277–303.

Tulving, E. (1972) "Episodic and Semantic Memory," in E. Tulving and W. Donaldson, eds, 'Organization of Memory,' Academic press, NY and London, pp. 381–402.

Wirth, R. (2010) “HB-CBC, HB-Best-Worst_CBC or NO HB at All,” *Sawtooth Software Conference Proceedings*, 321–356.