Comparing Apples to Oranges

To produce useful insights, between-subject comparisons must have a common origin.

By Lynd Bacon, Peter Lenk, Katya Seryakova, and Ellen Veccia

Jane, the marketing research director for Acme Litter Box Manufacturing Co., was asked by Acme's vice president of marketing to investigate customers' preferences for the new eco-friendly, power-ventilated litter box the company had in the planning stage. The VP wanted to better understand customers' issues regarding house cat management. She wanted to dig deep for better litter intelligence.

Jane fielded a preliminary quantitative survey, followed by some focused qualitative research on specific issues that might be uncovered. She based her survey on previous results and intended to assess cat owners' concerns, including litter attitudes and usage. Jane also measured the importance of different cat maintenance issues, like bad breath and furniture scratching, to customers. In the hope of being able to discriminate well between items and customers, she used MaxDiff scaling, a choice-based method, rather than importance rating scales.

Executive Summary

MaxDiff scaling and discrete-choice conjoint methods measure subjects' preference structures relevant to a common referent, thus removing a common scale origin for between-subject comparisons. Subject-level estimates of the partworths from discrete-choice and MaxDiff are not on a common scale across subjects; they do not have the same scale "zero point." This article describes augmenting choice-based tasks with ratings to recover the lost origin. The benefit is to extend the utility of discrete-choice methods to domains such as segmentation and targeting.

MaxDiff scaling, also known as best-worst analysis, is a type of discrete-choice exercise that can be used to scale items like product features or benefits on a single, underlying dimension such as importance or preference. Adam Finn and Jordan Louviere described this technique in a seminal 1992 article about public safety concerns ("Determining the appropriate response to evidence of public concern: The case of food safety," *Journal of Public Policy & Marketing*, 11 (1), 12-25).

Since then, MaxDiff has become a popular alternative to ratings and rankings. For instance, in 2003 Cohen and Neira described using it for segmentation in cross-country studies ("Measuring preference for product benefits across countries: Overcoming scale usage bias with Maximum Difference Scaling." ESOMAR 2003 Latin America Conference Proceedings. ESOMAR: Amsterdam, The Netherlands). In 2007, Flynn, Louviere, Peters, and Coast summarized its use in health care ("Best-worst scaling: What it can do for health care research and how to do it," *Journal of Health Economics*, 26, 171-189).

MaxDiff questions are based on an experimental plan, like conjoint profiles are. Each question or task presents a set of items, and the survey taker is asked to pick the "best" or "most" item, depending on the specific application, and also the "worst" or "least" item, from each set. Because MaxDiff is often used to measure preference or importance, the questions asked about each set of items are "Which of the following product features is the most important to you? Which is the least important?" or "Which do you most prefer?" and so on. Many researchers believe that MaxDiff tasks produce better data than other methods. MaxDiff discriminates between items being evaluated better than ratings and produces nearly twice the information at around the same cost as pick-best tasks.

Jane fielded her study and collected survey responses from litter box customers. Her MaxDiff exercise included tasks that had four concerns each. Jane analyzed her MaxDiff data using Sawtooth Software's HB-CBC, which implements an hierarchical Bayes (HB) multinomial logistic regression model to estimate individual-level importance coefficients, or partworths.

Jane then used her estimates to cluster her survey takers into two groups based on what was most important to them. She created a group primarily concerned about litter-related issues, such as litter dust and smell and a group irritated about their cats' impact on their domicile environment, such as cat dander, eating houseplants, and furniture scratching.

Acme's marketing VP wanted to understand these groups better, and so Jane invited some of the customers who took her survey to focus groups. Jane scheduled groups for the "litter issues" customers. It became apparent during the first focus group that the customers were not in agreement about the importance of litter issues. Some felt that litter issues were very important, others a lot less so. Of the latter, one commented, "In the bigger scheme of things, the litter box isn't really that important to me."

Jane and the VP were perturbed by the apparent lack of consensus in the litter-issue group, but decided it must be a fluke and went on to run some domicile-environment groups. Much to their chagrin, the same thing occurred: There was substantial diversity of opinions about the importance of domicile-related concerns. As this became disturbingly apparent, Jane and the VP passed a note to the moderator to ask the participants about what they said when they took the survey. The moderator reviewed the MaxDiff survey questions with the group participants. She then went around the table, asking each person about the MaxDiff items and the opinions that they expressed in the session. One group participant said: "Plant eating is important to me, compared to the other things you asked me about in the survey. Overall, it's not all that important to me. It's just more important than the other things you showed me."

"How could that have happened?" Jane asked. "The survey seems solid and the subjects were well-screened. We have obtained good market share predictions in other HB, discrete-choice, and MaxDiff studies. Why did our segmentation not align with the subjects' concerns?"

Blame It on Ipsativity

How did Jane and Acme go wrong? It could have been many things. For example, the customers participating in the research may not have had well-formed, stable preferences. But a very plausible reason for the disappointing results is that they were comparing and making decisions about customers based on information they didn't have.

Discrete-choice data, including MaxDiff data, are *ipsative*: They measure relative, rather than absolute, preferences or importance. Results based on ipsative scales are only comparable within a respondent, and not across respondents. When using partworth estimates of item importance or preference, you can sort items based on importance or preference within respondents, but sorting respondents based on their partworths does not give the results that you might expect. When two subjects give the same rank to an item, it does not then follow that the item is equally valuable or important to them.

Subject-level estimates of the partworths from discrete-choice, MaxDiff, and paired comparison studies are not on a common scale across subjects. They do not have the same

scale "zero point." As a result, doing things with ipsative scales, such as segmentation analysis and targeting, may not be very useful. It could also lead you very much astray, like it did Jane and Acme.

Exhibit 1 illustrates the effect of not having a common scale origin when using ipsative data. It shows a stylized situation with four customers—Jim, Maya, Mary, and Jong—and four litter box concerns—A, B, C, and D. In panel A (top), the importance evaluations are on a common scale that allows between-person comparisons. Jim feels the most intensely about the four concerns, and Maya feels the least intensely, overall. Jim and Jong give the same importance to A, while Jim views C as more important than Jong does.

Ipsative scaling removes the common origin because one of the options is selected as the base or reference option for estimation purposes. This option is arbitrarily given the value of zero. Panel B of Exhibit 1 illustrates the situation where concern D is arbitrarily assigned to be the base option. Then the measured importance of the other options is relative to D. Even though Jim and Jong have given A the same importance for option A on the absolute scale in panel A, it appears as though Jong rates A much higher than Jim does on the relative scales. The correct interpretation is that the difference in importance between A and D is greater for Jong than Jim.

Comparing Results

The basic problem is that there is insufficient information in discrete-choice responses to estimate a scale origin common to all respondents in a sample. One kind of solution augments the discrete-choice data with "auxiliary" data that provides a common origin. Ulf Böckenholt of McGill University addressed this as well as other solutions in the context of paired comparison data. In a 2004 article, he considered three different approaches ("Comparative judgments as an alternative to ratings: Identifying the scale origin," *Psychological Methods*, 9 (4), 453-465). One solution is to assume away the problem by a priori specifying a common origin. If the researcher knew each subject's importance rating for option D in panel B of Exhibit 1, then she could reconstruct panel A. Obviously, this approach has limited applicability in most marketing research contexts.

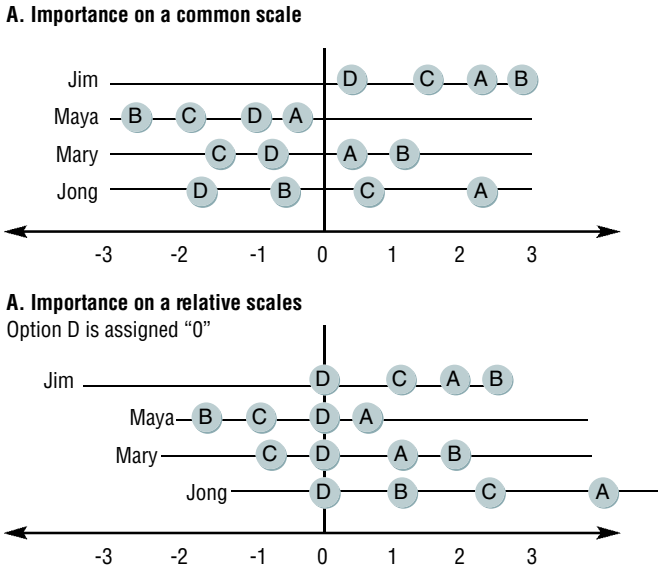
A second method Böckenholt described compares individual items and "bundles" consisting of combination of the individual items. A critical assumption is that a subject's overall evaluation of a bundle is the sum of its constituent parts. For example, the importance of the bundle with options A and D is the sum of the importance of A and D evaluated separately. Consider, for example, the following paired comparison question about less than happy outcomes.

Which would you most prefer? Assume that two cats are the same on all attributes other than the ones described: a cat that bites or a cat that sheds heavily and eats plants. By cleverly arranging the comparisons and using the appropriate contrasts, it would be possible to infer the absolute importance of all of the options across subjects.

Böckenholt's third method, which we advocate, combines relative and absolute judgments, for example, using data from both a discrete-choice exercise and importance ratings. In our approach to this scale origin problem, we use a model that fuses ratings and discrete-choice data. We model all responses using contemporary Bayesian statistical methods that allow for both scale-usage heterogeneity among subjects and method biases from using two elicitation procedures. We describe our modeling approach here in summary form.

Complete technical details are provided in a technical paper that is available from us upon request. A paper we presented at the 2007 Sawtooth Software Conference Proceedings in Sequim, Wash. ("Making MaxDiff more Informative: Statistical Data Fusion by way of Latent Variable Modeling.") provides additional background on related applications, as well as comparisons of different modeling approaches.

Exhibit 1 Uncorrelated variables



Estimating a Common Scale Origin

Exhibit 2 summarizes our modeling approach. The partworths or preferences are the main quantities of interest in the model. The model allows these subject-level parameters to vary across the sample where their heterogeneity can be related to subject-level covariates. The covariates are in dotted ovals to show that they are optional model components. The partworths then drive two submodels. The submodel on the left in Exhibit 2 describes the discrete-choice responses, and the one on the right is for our auxiliary importance ratings data. The discrete-choice submodel is similar to the standard one based on the traditional, random utility model described by Daniel McFadden in his Nobel Prize-winning work, "Conditional logit analysis of quantitative choice behavior," (Frontiers in Econometrics, ed by P Zarembka, Academic Press, New York, 105-142).

The diagram indicates that the subject combines his or her partworths with random error to arrive at the observed choices. For a "pick-the-most-preferred" type of study, like

conventional choice-based conjoint, the most preferred option is linked to the maximum utility, while for MaxDiff studies the most and least preferred options are linked to the maximum and minimum utilities of the choice sets.

Our choice model differs from standard models in one important aspect: We do not set one of the partworths equal to zero. When fitting choice data alone, not all of the partworths can be uniquely estimated, and constraints are needed to identify the model. As a result, the relative difference between each partworth and a base option is estimated, thus removing the common origin for between-subject comparisons. This compromises the efficacy of choice-based studies for segmentation and targeting.

To avoid needing to estimate differences between partworths, we augment the discrete-choice data with ratings data. The ratings data provide information about overall level differences between subjects and allows us to estimate a common origin for the partworths. We use what is often called a "cut-point model" to relate the observed, ordinal ratings to continuous, latent variables. The cut-point model accommodates scale-usage heterogeneity. These latent variables, in turn, are driven by the partworths.

We identify the partworths by moving the model identification constraints from the discrete-choice model to the ratings model. Consequently, the partworths in the joint model are not constrained at all. An important side benefit of our model is that it allows us to test whether the discrete-choice and ratings tasks tap into the same or different cognitive processes. It's possible that subjects in any given study may rely on different choice processes when responding to different tasks. The ratings submodel also moderates the partworths to ameliorate well-known method biases.

In theory, one ratings item in the survey is sufficient to recover the common origin, provided that it depends on one or more of the partworths. In practice, multiple ratings should provide a more robust estimate of the common origin. An unexplored question is the number, type of items, and ratings scale that best facilitate inter-subject comparisons in any given context.

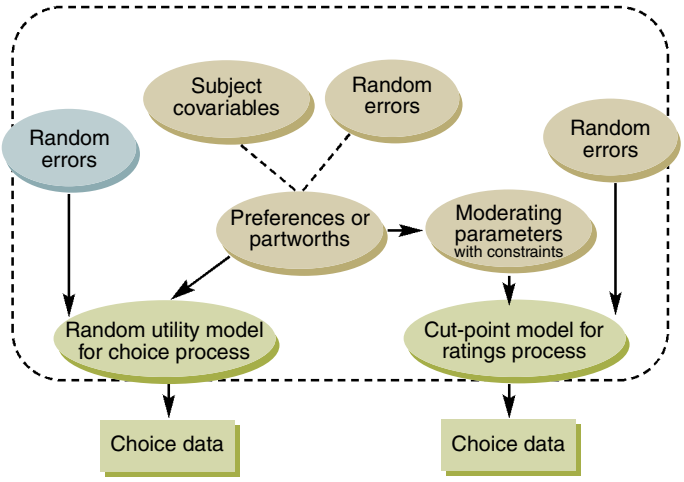
We estimate the joint model for discrete-choice and ratings data simultaneously using Markov chain Monte Carlo methods. A complete technical description of our modeling approach is in a working paper we will be happy to provide upon request.

Example: Cat Owners' Concerns

Jane contracted a marketing research consultant to field a survey that probed into cat owner's issues, while also collecting enough information to estimate a common scale origin. There were 12 issues considered, identified for brevity as C1 to C12.

C1 Shedding
C2 "Singing" at night
C3 Litter box smell
C4 Furniture scratching
C5 Finicky eater
C6 Litter box dust
C7 Biting
C8 Plant eating
C9 Bad breath
C10 Dry skin
C11 Good fur
C12 Healthy and happy

Exhibit 2 Choice model with common origin



Acme's questionnaire included a MaxDiff task, which consisted of selecting the most and least important concerns. Before the MaxDiff task, the subjects rated each of the 12 concerns on a 1 to 5 scale from "Not at All Important" to "Very Important." Acme's research company collected completed surveys from 300 customers.

Exhibit 3 summarizes the MaxDiff data from the survey. Using Acme's data we fitted a standard hierarchical Bayes model to the MaxDiff data without the ratings data and our proposed model that fuses both MaxDiff and ratings data. In the standard model, C12 was the reference concern. It was assigned a partworth of zero. The partworths for C1 to C11 were estimated within each subject relative to C12. Our model freely estimated the partworths for all 12 concerns and preserved the common origin for making inter-subject comparisons.

Concerns are on the x-axis, sorted in ascending order of importance. The y-axis to the left indicates response frequency, and the y-axis on the right indicates rating on the five-point scale used. The bars indicate the frequency of most and least choices aggregated over the sample. The line graph indicates average importance.

To illustrate the importance of having a common origin, we sorted the surveyed customers based on the sum of their estimated partworths. We then put them into three tiers where the first tier consists of 100 customers with the lowest summed partworths, the second tier is the next 100, and the third tier is the 100 customers with the largest partworth sums. This third tier consists of subjects who have the largest overall concern about cat ownership issues. The first tier has the lowest overall level of concern.

Exhibit 3 Summary measures of cat owners' concerns

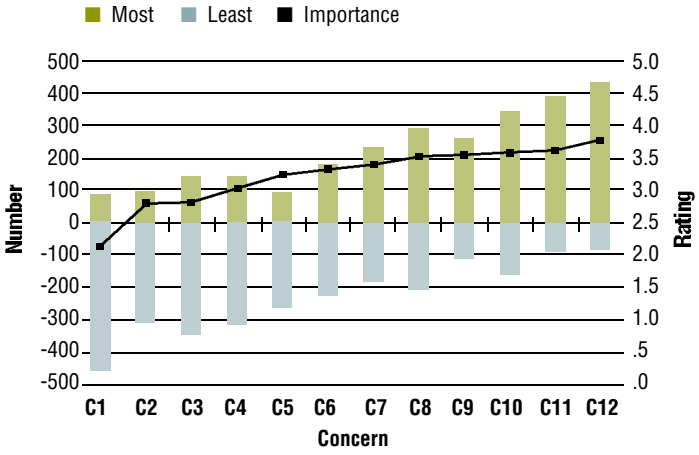


Exhibit 4 Box plots showing distribution of estimated partworths

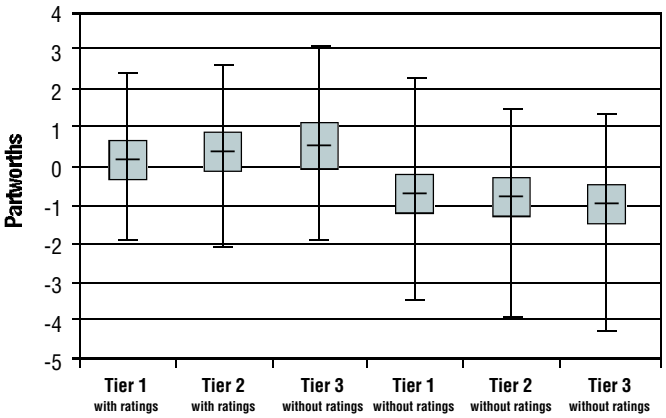


Exhibit 4 shows box plots for the partworths in each of the three tiers for the two models. The three box plots on the left are the tiers based on our model and with a common origin, while the three on the right are from the standard MaxDiff model. The distributions with a common origin exhibit an upward trend in preferences. The distributions without a common origin, on the other hand, do not.

The box plots show distribution of estimated partworths as a function of importance tier and whether a common origin was estimated using ratings (the three box plots on the left) or not (three box plots on the right). The line in the middle of the box is the median; the top of the box is the 75th percentile, and the bottom of the box is the 25th percentile. The "whiskers" extending from the box indicate the range of the data.

Acme's survey also collected attitudinal data about cats and the owners' lifestyles. The responses were used to construct two multi-item scales: Cat Devotion, where higher scores indicate more concern about their cats, and Domicile Devotion, where higher scores indicate relatively high concern about their domicile environment. We added these scales to the model in Exhibit 2 as subject-level covariates that drive subject-level preference differences.

The estimated regression coefficients in Exhibit 5 indicate that subjects that scored higher on Cat Devotion also tend to view C5, C7, C8, C10, C11, and C12 as more important issues. Subjects that scored higher on Domicile Devotion tend to be more concerned about C3, C4, and C6 and relatively less concerned about C9 and C10. As it turned out, C10 was shown to be a pivotal concern that is positively related to Cat Devotion and negatively related to Domicile Devotion. Jane also scheduled focus groups with subjects from these two groups and, unlike her first study, confirmed the segmentation and targeting of the second study. To make a long story short, using this information, Jane designed the positioning, packaging, and communications for their eco-litter box and achieved a major sales success.

A Clearer Picture

Standard, discrete-choice models measure preferences or importance within a subject relative to a base alternative. These ipsative scales suffice for some marketing activities, such as market share simulators or sorting items within subjects. However, the loss of a common origin impedes between-subject comparisons, which are needed for effective segmentation and targeting. In this article, we propose a modeling approach that integrates ratings and discrete-choice data so that overall differences between respondents are not lost.

One question you may be asking yourself is whether all this machinery is really necessary. As is usually the case, it depends on the study's objectives. If your goal is product-line optimization of preference shares, then standard procedures should work very well because preference shares only depend on within-subject, relative preferences. If, on the other hand, you also desire to identify subjects according to their absolute preference, such as in segmentation and targeting, then the standard methods distort the comparisons because the preferences

Exhibit 5 Estimated coefficients between concerns and attitudinal factor

Concern devotion	Cat devotion	Home	Concern devotion	Cat devotion	Home
C1	0.085	0.003	C7	0.155	-0.059
C2	0.042	-0.016	C8	0.251	-0.064
C3	-0.069	0.417	C9	0.064	-0.306
C4	-0.178	0.144	C10	0.338	-0.263
C5	0.183	0.070	C11	0.170	-0.023
C6	-0.001	0.213	C12	0.299	-0.171

are on relative and not absolute scales. The proposed method provides a clearer picture of the subjects' intensity of preferences by allowing for between-subject comparisons.

Both approaches give the same preference orderings within subjects, but the new method we have described here extends the domain of application for discrete-choice methods. Many a study begins with the goal of finding product attributes or concerns that affect market shares, only to have the client make the additional request for segmentation or targeting after the data have been collected. Jane's post-processing of the relative partworths is the expedient, common approach. However, by including rating questions, such as importance or likelihood of purchase, which can often be collected with MaxDiff or discrete-choice tasks, the researcher is able to expand her analysis to better meet client needs.

Finally, it's worth noting that our general approach can use auxiliary data other than ratings. The data could be purchase quantity data instead, for example, so long as the auxiliary data can be linked to specific subjects in the study. What we have described here is an example of a general approach to obtaining more useful insights for marketing decision-making by using models to combine different kinds of data. We believe that this kind of model-based data integration will become increasingly common as more varied data on individuals becomes available. ●

Additional Reading

Cohen, S. and B. Orme (2004), "What's Your Preference?" *Marketing Research*, Summer, 32-37.

Lenk P. and L Bacon (2007), "Estimating Common Utility Origins and Scales in Discrete-Choice Conjoint with Auxiliary Data," working paper, The University of Michigan.

Lenk, P., W. DeSarbo, P. Green, and M. Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15 (2), 173-191.

Orme, B. (2005). *Getting Started with Conjoint Analysis*, Research Publishers, Madison, WI.

Lynd Bacon is president of Loma Buena Associates (www.LBA.com). He may be reached at lbacon@LBA.com. Peter Lenk is associate professor, Stephen M. Ross Business School, the University of Michigan. He may be reached at plenk@bus.umich.edu. Katya Seryakova is senior project director, advanced analytics and consulting, Knowledge Networks, Inc. She may be reached at kseryakova@knowledgenetworks.com. Ellen Veccia is senior vice President of the advanced analytics group at Knowledge Networks, Inc. She may be reached at eveccia@knowledgenetworks.com.