Winner of the John and Mary Goodyear Award for the Best International Paper presented at ESOMAR Conferences in 2003.

MEASURING PREFERENCE FOR PRODUCT BENEFITS ACROSS COUNTRIES

Overcoming scale usage bias with Maximum Difference Scaling

Steve Cohen
Leopoldo Neira

INTRODUCTION

The measurement of consumer preferences has long been an area of interest to both academic and practicing researchers. Accurate measurement of preferences allows the marketer to gain a deeper understanding of consumers' wishes, desires, likes, and dislikes, and thus permits a better implementation of the tools of the marketer. After measuring preferences, a common activity is market segmentation, which permits an even more focused execution of the marketing mix.

Since the mid-1950s, marketing researchers have responded to the needs of management by conducting market segmentation studies. These studies are characterized by the collection of descriptive information about benefits sought, attitudes and beliefs about the category, purchase volume, buying styles, channels used, self, family, or company demographics, and so on. Upon analysis, the researcher typically chooses to look at the data through the lens of a segmentation basis. This basis is either defined by preexisting groups – like heavy, medium, and light buyers or older vs. younger consumers – or defined by hidden groups uncovered during an in-depth statistical analysis of the data – benefits segments, attitude segments, or psychographic segments. Finally, the segments are then cross-tabulated against the remaining questions in the study to profile each group and to discover what characteristics besides the segmentation base distinguish them from one another.

Quite often, researchers find that preexisting groups, when different, are well distinguished in obvious ways and not much else. Wealthier consumers buy more goods and services, women buy and use products in particular categories more than men, smaller companies purchase less and less often than larger companies, and so on. However, when looking at buying motivations, benefits sought, and their sensitivity to the tools of marketers (e.g. price, promotions, and channel strategies), members of preexisting groups are often found to be indistinguishable from one another.

This realization has forced researchers to look to *post hoc* segments formed by a multivariate analysis of benefits, attitudes, or the like. This focus on benefits, psychographics, needs and wants, and marketing elasticities as means of segmentation has gained favor since the early work of Haley (1985) and is the mainstay of many market segmentation studies currently conducted. The utility of a focus on *post hoc* methods has been widely endorsed by marketing strategists (Aaker, 2001):

> *"If there is a 'most useful' segmentation variable, it would be benefits sought from a product, because the selection of benefits can determine a total business strategy."*

Using Benefits Segmentation as our example, we compare two methods of measuring preferences for benefits and then extracting benefits segments. One method uses the traditional method of rating benefits on 5-point scales and then using Cluster Analysis to extract the segments. We also examine a newly introduced method, called Maximum Difference Scaling (*MaxDiff*), to provide the benefits ratings, and then we use a Latent Class Model to extract the segments.

This paper is organized as follows. We first briefly review the standard practices of benefit measurement and benefit segmentation and, along the way, point out their deficiencies. We then introduce the reader to Maximum Difference Scaling, a method that we believe is a much more powerful method for measuring benefit importance – a method that is *scale-free* and thus very applicable to international segmentation research. The next section describes how Maximum Difference Scaling can be combined with Latent Class

Analysis to obtain international benefit segments. We then describe an example of how both the traditional and the newer methods were used in a cross-national consumer study of coffee drinkers and we compare the results. We conclude with some final thoughts and suggestions for use.

TRADITIONAL SEGMENTATION TOOLS

The two-stage or "tandem" segmentation method has been used for over twenty years (Haley, 1985), and has been described by Myers (1996) as follows:

1. Administer a battery of rating-scale items to a group of consumers, buyers, customers, etc. These rating scales typically take the form of agree/disagree, describes/does not describe, important/not important ratings. Scales of five, seven, ten, or even 100 points are used.
2. The analyst then seeks to reduce the data to a smaller number of underlying dimensions or themes. Factor Analysis of the rating scale data, using either the raw ratings or some transformation of the ratings (like standardization) to obtain better statistical properties, is most often performed. The analyst then outputs the factor scores, one set of scores for each respondent.
3. The factor scores are passed to a Cluster Analysis, with k-means Cluster Analysis being the most preferred and the most often recommended by academic researchers (Punj and Stewart (1983). K-means is implemented in SAS as Proc Fastclus and in SPSS as Quickcluster.
4. The clusters are profiled. A cross-tabulation of group, cluster, or segment membership is created against all the other significant items in the survey.

This tandem technique of market segmentation has many perils, among them are scalar equivalence and the continued use of Factor Analysis followed by Cluster Analysis for segmentation.

Scalar equivalence (Steenkamp and ter Hofstede, 2002) refers to the fact that, even though the segmentation basis variable may have been operationalized to be the same across countries, the scores obtained may not be directly comparable across countries. One of the common causes of scalar inequivalence is country-level response styles, where a response style is the tendency of people within a country to respond systematically to items based on some other criterion than what is being measured. The best-known response styles are acquiescence bias, extreme responding, and social desirability (Paulhus, 1991). There is ample evidence (Chen, Lee, and Stevenson, 1995; Steenkamp and Baumgartner, 1998; ter Hofstede, Steenkamp, and Wedel, 1999; Baumgartner and Steenkamp, 2001) that countries differ in their response styles. We note that scalar inequivalence is *less likely* to occur when collecting constant sum or ranking data. Constant sum data forces trade-offs and avoids yea saying. However, constant sum data may be difficult to collect if there are many items. Another alternative may be ranking the benefits. However, the major advantage of ranking – each scale point is used once and only once – may be outweighed by the fact that ranking suffers from order effects, does not allow ties, and is not appropriate when absolute scores are needed (e.g. purchase intent ratings).

Hence, we conclude that we would like a rating method that does not experience scale use bias, forces trade-offs, and allows each scale point to be used once and only once.

For grouping people, the tandem method of segmenting respondents using factor scores followed by Cluster Analysis is a very common practice. Cluster Analysis may be characterized as a *heuristic* method since there is no *underlying* model being fit. We contend that, while using Factor Analysis gets rid of the problems associated with correlated items, it introduces the problems of which factoring method to use, what type of rotation to use, factor score indeterminacy, and the selection of the final number of factors.

Deriving patterns from Factor Analysis and making cross-country comparisons becomes problematic when ratings have systematic scale use biases and item inter-correlations. For example, when using a rating scale in a segmentation analysis, the first dimension uncovered in a Factor Analysis often tends to be a general factor. Using this factor in a Cluster Analysis will often uncover a "high rater" segment or a "general" segment. Additional partitions of the data may uncover meaningful groups who have different needs, but only after separating out a group or two defined by their response patterns. This approach is especially dangerous in multi-country studies, where segments often break out on national lines, more often due to cultural differences in scale use than to true differences in needs. Indeed, as noted by Steenkamp and ter Hofstede (2002):

> *"Notwithstanding the evidence on the biasing effects of cross-national differences in response tendencies, and of the potential lack of scalar equivalence in general on the segmentation basis, it is worrisome to note that this issue has not received much attention in international segmentation research. We believe that cross-national differences in stylistic responding is one of the reasons why international segmentation studies often report a heavy country influence."*

Using Cluster Analysis alone has a number of limitations. These include forcing a deterministic classification (each person belongs absolutely to one and only one segment) and poor performance when the input data are correlated. In such situations, highly correlated items are “double counted” when perhaps they should be counted only once.

Academic research has rightly pointed out the deficiencies of the two-stage or tandem approach of Factor Analysis followed by Cluster Analysis (see DeSarbo et al, 1990; Dillon, Mulani, and Frederick, 1989; Green and Krieger, 1995; Wedel and Kamakura, 1999; and Arabie and Hubert, 1994). While the frequent use of the tandem method is unmistakable because of its ease of implementation with off-the-shelf software, most practicing researchers have simply failed to hear or heed these warnings. The bluntest assessment of the weakness of the tandem method may be attributed to Arabie and Hubert (1994):

> *“Tandem clustering is an out-moded and statistically insupportable practice.”* (italics in original)

While Chrzan and Elder (1999) discuss possible solutions to the tandem problem, they conclude that a heavy dose of clustering and factoring must be done before engaging in the final segmentation analysis, which may use all or a selection of the raw variables, or may use the tandem method, depending upon the items, their intercorrelations, and other characteristics of the data.

The next sections describe the use of Maximum Difference Scaling instead of rating scales to measure the *relative importance* of benefits and of Latent Class Analysis, instead of Cluster Analysis, as a method for uncovering market segments with similar benefit importances.

MAXIMUM DIFFERENCE SCALING

Maximum Difference Scaling (*MaxDiff*) is a measurement and scaling technique originally developed by Jordan Louviere and his colleagues (Louviere, 1991, 1992; Louviere, Finn, and Timmermans, 1994; Finn and Louviere, 1995; Louviere, Swait, and Anderson, 1995; McIntosh and Louviere, 2002). Most of the prior applications of MaxDiff have been for use in Best-Worst Conjoint Analysis. In applying MaxDiff to B-W Conjoint, the respondent is presented with a full product or service profile as in traditional Conjoint. Then, rather than giving an overall evaluation of the profile, the respondent is asked to choose the attribute/level combination shown that is most appealing (best) and least appealing (worst).

We apply this scaling technique instead to the measurement of the importance of product benefits and uncovering segments. This discussion follows the one made by Cohen and Markowitz (2002).

MaxDiff finds its genesis in a little-investigated deficiency of Conjoint Analysis. As discussed by Lynch (1985), additive conjoint models do not permit the separation of importance or weight and the scale value. Put another way, Conjoint Analysis permits *intra-attribute* comparisons of levels, but does not permit *across attribute* comparisons. This is because the scaling of the attributes is unique to each attribute, rather than being a method of global scaling.

Maximum Difference Scaling permits intra- and inter-item comparison of levels by measuring attribute level utilities on a common, interval scale. Louviere, Swait, and Anderson (1995) and McIntosh and Louviere (2002) present the basics of MaxDiff, or Best-Worst scaling. To implement Maximum Difference Scaling for benefits requires these steps.

Select a set of benefits to be investigated.

- Place the benefits into several smaller subsets using an experimental design (e.g. 2^k, BIB, or PBIB are most common). Typically over a dozen such sets of three to six benefits each are needed, but each application is different.
- Present the sets one at a time to respondents. In each set, the respondent chooses the most salient or important attribute (the best) and the least important (the worst). This best-worst pair is *the pair* in that set that has the Maximum Difference.
- Since the data are simple choices, analyze the data with a multinomial logit (MNL) or probit model. An aggregate level model will produce a total sample benefit ordering.
- Analyze pre-existing subgroups with the same statistical technique.
- To find benefit segments, use a Latent Class multinomial logit model (see below for discussion).

The MaxDiff model assumes that respondents behave *as if* they are examining every possible pair in each subset, and then they choose the most distinct pair as the best-worst, most-least, *maximum difference* pair.

Properly designed, MaxDiff will require respondents to make trade-offs among benefits. By doing so, we do not permit anyone to like or dislike all benefits. By definition, we force the relative importances out of the respondent. A well-designed task will control for order effects. Each respondent will see each item in the first, second, third, etc. position across benefit subsets. The design will also control for context effects: each item will be seen with every other item an equal number of times.

The MaxDiff procedure will produce a unidimensional interval-level scale of benefit importance based on nominal level choice data. As such, MaxDiff is particularly valuable in international segmentation research. Because there is only one way to choose something as "most important," there is no opportunity whatsoever to encounter bias in the use of a rating scale. Hence, there is no opportunity to be a constant high/low rater or a middle-of-the-roader. The method forces respondents to make a discriminating choice among the benefits. Looking back to the observations by Steenkamp and ter Hofstede, we believe that this method overcomes very well the problems encountered in cross-national attribute comparisons that are due to differences in the use of rating scales across countries. The MaxDiff method is easy to complete (respondents make two choices per set), may also control for potential order or context biases, and is rating scale-free.

LATENT CLASS ANALYSIS

We use the data from the MaxDiff task in a Latent Class (finite mixture) choice model (DeSarbo, Ramaswamy, and Cohen, 1995; Cohen and Ramaswamy, 1998) leading to easily identifiable segments with differing needs. All of this occurs in a scale-free and statistical-model-based environment. For readers not familiar with Latent Class Analysis, we present this short description of its advantages. Interested readers are referred to Wedel and Kamakura (1999) for a more detailed discussion.

Latent Class Analysis (LCA) has a great deal in common with traditional Cluster Analysis, namely the extraction of several relatively homogeneous and yet separate groups of respondents from a heterogeneous set of data. What sets LCA apart from Cluster Analysis is its ability to accommodate both categorical and continuous data, as well as descriptive or predictive models, all in a common framework. Unlike Cluster Analysis, which is data-driven and model-free, LCA is model-based, true to the measurement level of the data, and can yield results which are stronger in the explanation of buyer behavior.

The major advantages of LCA include:

- Conversion of the data to a metric scale for distances is not necessary. LCA uses the data at their original level of measurement.
- LCAs can easily handle models with items at mixed levels of measurement. In Cluster Analysis, all data must be metric.
- LCA fits a statistical model to the data, allowing the use of tests and heuristics for model fit. The tandem method, in contrast, has two objectives, which may contradict one another: factor the items, then group the people.

- LCA can handle easily cases with missing data.
- Diagnostic information from LCA will tell you if you have overfit the data with your segmentation model. No such diagnostics exist in Cluster Analysis.
- Respondents are assigned to segments with a probability of membership, rather than with certainty as in Cluster Analysis. This allows further assessment of model fit and the identification of outliers or troublesome respondents.

Perhaps the biggest difference between Cluster Analysis and LCA is the types of problems they can be applied to. Cluster Analysis is solely a descriptive methodology. There is no independent-dependent, or predictor-outcome relationship assumed in the analysis. Thus, while LCA can also be used for descriptive segmentation, its big advantage lies in simultaneous segmentation and prediction.

If we think of a discrete choice model as a predictor-outcome relationship, then we can apply an LCA. In this case, the outcomes or response variables are the Most and Least choices from each set and the predictors are the relative importance of the items being shown in each choice set. Recognizing the need for conducting *post hoc* market segmentation with Choice-based Conjoint Analysis (CBCA), DeSarbo, Ramaswamy, and Cohen (1995) combined LCA with CBCA to introduce Latent Class CBCA, which permits the estimation of benefit segments with CBCA. LC-CBCA has been implemented commercially in a program from Sawtooth Software.

To summarize this and the prior section:

- We advocate the use of Maximum Difference scaling to obtain a unidimensional interval-level scale of benefit importance. The task is easy to implement, easily understood by respondents and managers alike, and travels well across countries.
- To obtain benefit segments from these data, we advocate the use of Latent Class Analysis. LCA has numerous advantages over Cluster Analysis, the chief among them being that it will group people based on their pattern of nominal-level choices in several sets, rather than by estimating distances between respondents in an unknown or fabricated metric.

The next section discusses an empirical example of the application of these techniques and compares them to results from using traditional tandem-based segmentation tools.

AN EXAMPLE

Coffee is a beverage that is enjoyed around the world. Our client, a multinational company offering coffee in many countries, wished to conduct a multi-country segmentation study of coffee drinkers. The goal of the research was twofold:

- Identify segments of consumers with similar benefits that they seek from drinking coffee; and,
- Compare the results of the new method of MaxDiff with LC Models vs. the traditional tandem clustering methods.

Data were collected in six Central American countries from coffee users in all socioeconomic strata. The principal grocery shopper over 18 years old in the household was interviewed. Upon being screened to be eligible for the study, approximately one-half of the consumers evaluated thirteen coffee benefits using a 5-point importance scale. The other half of the respondents were given thirteen sets of four benefits each, and were asked to choose the Most Important benefit and the Least Important benefit in each of the thirteen sets. In short, the first group provided thirteen ratings and the second group provided 26 (thirteen Mosts and thirteen Leasts) evaluations.

The percents by country and final sample size for the two methods are as follows (see table 1).

Table 1
PERCENT OF PEOPLE DOING EACH TASK BY COUNTRY

Country	*Ratings*	*Most-Least*
Panama	12%	14%
Costa Rica	17%	19%
Nicaragua	13%	13%
Honduras	23%	21%
El Salvador	25%	24%
Guatemala	10%	10%
Total N each task	709	703

The questionnaire contained the benefit ratings, basic demographics, and brand awareness and use.

Thirteen product benefits were provided by the client company. They are:

1. It always keeps me company
2. It makes me feel relaxed
3. It helps me to be sociable with others
4. It helps me to stay alert
5. It makes me feel good
6. It helps me to start my day with a positive attitude
7. It helps me to stay focused on my assignments
8. It has been present in the best moments of my life
9. It helps me to keep a good level of energy during the day
10. It helps my digestion
11. It helps me to keep warm
12. It makes me feel confident
13. It helps me to belong to different social groups

Since data collection was done as a paper and pencil exercise, the rating scale task kept the order of the items fixed, but started the ratings at a random start and then continued through the fixed list. Ratings of each benefit were asked on a five-point scale of Extremely Important, Very Important, Somewhat Important, Not Very Important, and Not at All Important.

To develop the MaxDiff task, we used an experimental design. We created thirteen sets of four attributes each. Across the sets, every possible pair of items appeared together exactly once. Each benefit appeared once in each of the four positions in a set (first, second, third, and fourth) and each benefit appeared exactly four times across the thirteen sets. When shown a set of four items, the respondents were asked to *choose the items that were the most important and the least important benefits when deciding to drink coffee*. The order of the thirteen sets was fixed for all respondents.

Benefit evaluations were asked first, brand awareness and use next, and then the demographic questions finished the survey.

Tandem Method Results

We first show the results of the ratings task. Table 2 shows top two box ratings for each of the items by country.

Table 2
Percent Extremely / Very Important by Country

	Total	*Panama*	*Costa Rica*	*Nica-ragua*	*Hon-duras*	*El Sal-vador*	*Guate-mala*
Has been present in the best moments of my life	57%	73%	39%	84%	62%	46%	53%
Helps my digestion	57%	79%	49%	78%	51%	50%	49%
Makes me feel good	54%	55%	27%	90%	60%	50%	53%
Helps me to start my day with a positive attitude	49%	50%	31%	78%	45%	51%	43%
Helps me to belong to different social groups	49%	46%	38%	49%	59%	49%	44%
Makes me feel confident	46%	60%	23%	58%	42%	58%	37%
Helps me to keep a good level of energy during the day	45%	48%	34%	73%	43%	43%	38%
Helps me to stay alert	45%	62%	25%	78%	38%	41%	40%
Helps me to keep warm	43%	45%	16%	68%	41%	55%	29%
Always keeps me company	42%	24%	34%	66%	38%	49%	43%
Helps me to be sociable with others	41%	26%	34%	69%	44%	37%	35%
Makes me feel relaxed	40%	39%	25%	63%	51%	30%	38%
Helps me to stay focused on assignments	39%	33%	23%	77%	32%	42%	37%

The items are shown in descending order of importance as based on top two box scores on the five-point scale. Note, in particular, the country differences in scale use. In particular, those respondents from Costa Rica tend to be lower rates than the total (all scores are below the total), those from Nicaragua are higher rates than the total (all but one item is rated higher than total), and the other countries show varying degrees of skew from the total, depending upon the item. The direct comparison of country scores is clearly violated because of these within-country tendencies.

We then subjected the full five point ratings to a Factor Analysis followed by a Cluster Analysis. The Factor Analysis derived the first eigenvalue as 5.1, the second as 1.4, the third as 0.9, and the fourth as 0.7. In passing, note that Chang (1983) and Dillon, Mulani, and Frederick (1989) point out that the factors with the largest eigenvalues are not necessarily the factors that best distinguish among people. While we could have used more than two factors in this analysis, we would have violated the common rule of only accepting eigenvalues greater than one. Thus a two factor solution is our choice. The Varimax Rotated Factor Pattern is shown in table 3.

The factor loadings are multiplied by 100 and rounded to the nearest whole number. The largest loadings on each factor are shaded. Despite our best efforts, we cannot find a managerially relevant interpretation of these factors.

Table 3
TWO FACTOR SOLUTION: PRINCIPAL COMPONENTS VARIMAX ROTATION

	Factor 1	*Factor 2*
It always keeps me company	11	**81**
It makes me feel relaxed	30	**64**
It helps me to be sociable with others	16	**80**
It helps me to stay alert	**71**	19
It makes me feel good	**71**	18
It helps me to start my day with a positive attitude	**65**	24
It helps me to stay focused on my assignments	**59**	24
It has been present in the best moments of my life	**66**	15
It helps me to keep good level of energy during day	17	**56**
It helps my digestion	**73**	7
It helps me to keep warm	**64**	36
It makes me feel confident	**64**	24
It helps me to belong to different social groups	39	44

Following this, we output the Factor Scores and used them in a k-means Clustering. Using SAS Fastclus (SAS Institute, 2002), we ran two to eight group cluster solutions. The Cubic Clustering Criterion was used to decide upon the "best" solution of four groups. Top two box importance scores for each cluster are shown in table 4.

Table 4
TOP TWO BOX SCORES BY SEGMENT

	Total	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 4*
Segment size	100%	37%	21%	29%	13%
It has been present in the best moments of my life	57%	81%	80%	26%	22%
It helps my digestion	57%	79%	85%	24%	21%
It makes me feel good	54%	82%	70%	21%	23%
It helps me to start my day with a positive attitude	49%	79%	54%	15%	24%
It helps me to belong to different social groups	49%	74%	45%	18%	46%
It makes me feel confident	46%	75%	57%	13%	18%
It helps me to keep a good level of energy during the day	45%	69%	34%	17%	57%
It helps me to stay alert	45%	70%	63%	11%	14%
It helps me to keep warm	43%	77%	48%	8%	12%
It always keeps me company	42%	80%	9%	7%	66%
It helps me to be sociable with others	41%	75%	15%	6%	59%
It makes me feel relaxed	40%	76%	21%	9%	37%
It helps me to stay focused on my assignments	39%	63%	53%	10%	13%

Notice that Cluster 1, the largest segment, consists of all high raters. Every single item is rated higher than by the total sample. Cluster 3 is a low rater segment: every rating is lower than the total sample. Cluster 4 tends to be low raters: 10 out of the 13 items are rated lower than the total and Cluster 2 is mixed. Seven of the items are rated high in Cluster 2, while six are rated low.

We then crosstabulated the segment memberships with country and found that there is a very strong association between country and cluster (Chi-square = 153.9, df = 15, p <.0001). As expected, the response biases evident in the clusters match up well with the countries of origin. For example, we observed earlier that Nicaragua as a whole were high raters; in fact, 60% of all Nicaragua respondents are in Cluster 1, the high rater segment.

Table 5
CLUSTER MEMBERSHIP WITHIN COUNTRY

	Total	*Panama*	*Costa Rica*	*Nica-ragua*	*Honduras*	*El Sa-vador*	*Guate-mala*
Cluster 1	37%	29%	11%	60%	39%	49%	32%
Cluster 2	21%	46%	15%	27%	14%	20%	13%
Cluster 3	29%	23%	45%	7%	31%	24%	40%
Cluster 4	13%	2%	30%	7%	16%	7%	15%

We thus conclude that the usual practices of tandem clustering have once again failed to distinguish segments of meaningful, managerially relevant significance, and instead display segments whose main differences may be attributed to differences in scale use. We now display the MaxDiff results accompanied by the Latent Class model results.

MAXDIFF AND LATENT CLASS RESULTS

To derive a utility or importance value for each benefit, the Most-Least choices are subjected first to an aggregate Multinomial Logit Model, and then to a Latent Class MNL model. The results are benefit utilities. In this study, the utilities for the benefits range in the aggregate from positive 1.8 to zero. We have found that looking at raw utilities may sometimes be unclear to managers. For ease of interpretation, we rescale the utilities according to the underlying choice model.

Remember that the model estimated is a multinomial logit (MNL) model, where the sum of the choices after exponentiating is 100%. Hence, if we rescale the utilities according to the MNL model, we will get a "share of preference" for each benefit. If all benefits were equally preferred, then each one's share of preference would be 7.7% (=1/13). If we index 7.7% to be 100, then a benefit with an index score of 200 would result from a share of preference of 15.4% (7.7% times 2). We have found that using this rescaling makes it much easier for managers and analysts to interpret the results. In this paper, we present only the index numbers and not the raw utilities.

By using the standard aggregate multinomial logit model, we obtained the results in table 6, after rescaling.

Table 6
TOTAL SAMPLE RESULTS FROM MAXDIFF TASK

	Total 100%
Makes me feel relaxed	177
Makes me feel good	171
Helps me to stay alert	134
Helps me to start my day with a positive attitude	122
Helps me to stay focused on my assignments	115
Has been present in the best moments of my life	101
Helps me to keep a good level of energy during the day	99
Always keeps me company	77
Helps me to keep warm	75
Helps my digestion	66
Makes me feel confident	65
Helps me to be sociable with others	57
Helps me to belong to different social groups	39

Note that the highest index numbers are very close for two items, *Makes me feel relaxed* and *Makes me feel good*, with both over 170. The next two are *Helps me to keep alert* (134) and *Helps me to start my day with a positive attitude* (122). These four benefits seem as if they should be at the top of the list and we deem the face validity of the results to be quite good. If we compare these results to the rating scales results, we find that *Makes me feel relaxed* is last in importance, and *Makes me feel good* is third. We contend that scale use bias and forcing no trade-offs may deliver these different results.

We then ran Latent Class MNL models (DeSarbo, Ramaswamy, and Cohen, 1995), generating two to seven Latent Classes. Using the usual information statistics (Kamakura and Wedel, 1999) to decide upon which solution to use, the AIC – the statistic which usually points to more latent classes, indicated that a seven group solution was best, while the CAIC – the statistic that usually points to fewer classes – recommended five classes. Upon further examination, the five class solution had segments whose sizes ranged from 25% to 12%, while the seven group solution had sizes ranging from 18% to 8%. We decided for reasons of parsimony, managerial use, and interpretability to use the five group solution.

The index numbers for this solution is displayed in table 7, along with the Total results for comparison purposes. The highest index numbers in each column are bold for ease of interpretation.

Table 7
RESCALED UTILITIES FOR 5 LATENT CLASSES

	Total	*Class1*	*Class2*	*Class3*	*Class4*	*Class5*
Segment Ns	703	178	154	162	125	84
Segment Size ------->	100%	25%	22%	23%	18%	12%
It makes me feel relaxed	**177**	77	**151**	**189**	**282**	**159**
It makes me feel good	**171**	92	**168**	148	**160**	**220**
It helps me to stay alert	134	77	**257**	125	61	89
It helps me to start my day with a positive attitude	122	117	140	**156**	82	20
It helps me to stay focused on my assignments	115	112	**250**	46	44	25
It has been present in the best moments of my life	101	116	61	39	77	**189**
It helps me to keep a good level of energy during the day	99	83	**152**	138	40	17
It always keeps me company	77	77	19	23	**352**	106
It helps me to keep warm	75	71	23	107	16	**361**
It helps my digestion	66	55	21	**245**	38	13
It makes me feel confident	65	71	31	58	44	32
It helps me be sociable with others	57	**152**	18	19	83	46
It helps me to belong to different social groups	39	**176**	10	8	21	24

The first segment (25% of the total sample) is interested in the *Sociability* that coffee brings. Note that none of the most important aggregate benefits are highly salient to this group.

- Members of the second segment (22%) drink coffee because it makes them focused and alert.
- The third segment (23%) likes how coffee *aids their digestion* and *helps them to be relaxed*.
- Those in the fourth segment (18%) think of *coffee as a friend*: something that keeps them company and helps them to be relaxed.
- The fifth and final segment (12%) is *warmed by coffee, makes them feel good*, and *has been associated with the best moments of their lives*.

These segments have a good deal of face validity and were meaningful and interpretable to our client.

We ran one final analysis (see table 8) to see if there was a large skew towards any country in the segment results. Table 8 shows some small skews by country, but they are barely statistically significant at the 95% confidence level (Chi-square = 31.7, df = 20, p = .047). We conclude that the MaxDiff method has not allowed response bias and that the Latent Class MNL model has produced benefit segments that are managerially interesting and relevant.

Table 8

Latent Class Membership within Country

	Total	***Panama***	***Costa Rica***	***Nicara-gua***	***Honduras***	***El Salvador***	***Guate-mala***
Class 1	25%	28%	25%	26%	23%	25%	27%
Class 2	22%	24%	21%	30%	19%	14%	36%
Class 3	23%	20%	26%	16%	24%	26%	21%
Class 4	18%	15%	18%	11%	19%	24%	12%
Class 5	12%	12%	11%	16%	14%	11%	4%

In a study that will soon be published in the Sawtooth Software Conference Proceedings, Cohen (2003) performed as similar comparison using a split sample of IT managers. In this case, ratings of the importance of 20 items were compared against 30 randomly-chosen paired comparisons of items and against a MaxDiff task composed of 15 sets of four items each. About 125 IT managers performed each task in a computerized interview over the Internet.

Several tests compared the results across methods. T-tests were conducted across the responses to judge the amount of discrimination each provided. The average t statistics of the ratings was 3.3, for paired comparisons it was 6.3, and for the MaxDiff results it was 7.7. This indicates that MaxDiff provides more discrimination within a person on his/her preferences than the other two methods.

A second test computed F-tests across 19 segmentation background variables. Since 20 items were rated, the 95% confidence level would expect to see 19 significant differences (20 items * 19 variables = 380 F-tests). For raw ratings, 30 of the F-tests were significant, 40 of the paired comparisons, and 38 of the tests involving MaxDiff.

Finally, a set of holdout ranking tasks was used to judge the predictive power of each method to correctly predict the ranks. For raw ratings, the percent correctly predicted was 85%, for pairs it was 88%, and for MaxDiff it was 97%. The conclusion of this study was that MaxDiff, in this dataset, was superior to either of the other two methods. Unfortunately, for this coffee study, we did not have the time to include such a holdout task,

SUMMARY

The intent of this paper has been to present practicing researchers with an innovative use of state-of-the-art tools to solve the problems that are produced when using traditional rating scales and the tandem method of clustering. We also compared the results of the suggested method against the traditional tools and found that the new tools provide “better” results.

Therefore, we suggest that practitioners adopt Maximum Difference scaling for developing a unidimensional scale of benefit importance. The MaxDiff task is easy for a respondent to do and it is scale-free, so that it can easily be used to compare results across countries. Furthermore, the tool is easy to implement, relatively easy to analyze with standard software, and easy to explain to respondents and managers alike.

To obtain benefit segments, we suggest using Latent Class Analysis. LCA has numerous advantages over Cluster Analysis. The disadvantages of this latter method are well known but not often heeded. The benefits of LCA have been demonstrated in many academic papers and books, so its use, while limited, is growing. We hope that this paper will spur the frequent use of these two methods.

This paper has shown that current research practice can be improved and that traditional methods are lacking and need to be updated. By describing the use of these newer methods and comparing them to traditional methods, we have shown that the modern researcher can overcome scale use bias with Maximum Difference Scaling and can overcome the many problems of Cluster Analysis by using Latent Class models.

References

Aaker, David A. (1995) *Strategic Market Management*. New York: John Wiley & Sons.

Arabie, Phipps and Lawrence Hubert (1994). Cluster analysis in marketing research. *Advanced Methods of Marketing Research*. Richard J. Bagozzi (Ed.). London: Blackwell Publishers, 160-189.

Baumgartner, Hans and Jan-Benedict E.M. Steenkamp (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, 38 (May).

Chang, Wei-Chen (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 32, 267-75.

Chen, C., S.Y. Lee, and H.W. Stevenson (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science* 6, 170-175.

Chrzan, Keith and Andrew Elder (1999). Knowing when to Factor: Simulating the tandem approach to Cluster Analysis. Paper presented at the Sawtooth Software Conference, La Jolla, CA.

Cohen, Steven H. (2003). Maximum Difference Scaling: Improved measures of importance and preference for segmentation. Paper presented at the Sawtooth Software Conference, San Antonio. TX.

Cohen, Steven H. and Venkatram Ramaswamy. (1998) Latent segmentation models. *Marketing Research Magazine*, Summer, 15-22.

Cohen, Steven H. and Paul Markowitz. (2002) Renewing Market Segmentation: Some new tools to correct old problems. *ESOMAR 2002 Congress Proceedings*, 595-612, ESOMAR: Amsterdam, The Netherlands.

DeSarbo, Wayne S., Kamel Jedidi, Karen Cool, and Dan Schendel. (1990) Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups. *Marketing Letters*, 3, 129-146.

DeSarbo, Wayne S., Venkatram Ramaswamy, and Steven H. Cohen. (1995) Market segmentation with choice-based conjoint analysis. *Marketing Letters*, 6 (2), 137-47.

Dillon, William R., Narendra Mulani, and Donald G., Frederick. (1989) On the use of component scores in the presence of group structure. *Journal of Consumer Research*, 16, 106-112.

Finn, Adam and Jordan J. Louviere. (1992) Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy and Marketing*, 11:1, 19-25.

Green, Paul E. and Abba Krieger. (1995) Alternative approaches to cluster-based market segmentation. *Journal of the Market Research Society,* 37 (3), 231-239.

Haley, Russell I. (1985) *Developing effective communications strategy: A benefit segmentation approach.* New York: John Wiley & Sons.

Louviere, Jordan J. (1991) Best-worst scaling: A model for the largest difference judgments. Working paper. University of Alberta.

Louviere, J.J. (1992). Maximum difference conjoint: Theory, methods and cross-task comparisons with ratings-based and yes/no full profile conjoint. Unpublished Paper, Department of Marketing, Eccles School of Business, University of Utah, Salt Lake City.

Louviere Jordan J., Adam Finn, and Harry G. Timmermans (1994). Retail Research Methods. *Handbook of Marketing Research*, 2nd Edition, McGraw-Hill, New York.

Louviere, Jordan J., Joffre Swait, and Donald Anderson. (1995) Best-worst Conjoint: A new preference elicitation method to simultaneously identify overall attribute importance and attribute level part worths. Working paper, University of Florida, Gainesville, FL.

Lynch, John G., Jr. (1985) Uniqueness issues in the decompositional modeling of multiattribute overall evaluations: An information integration perspective. *Journal of Marketing Research*, 22, 1-19.

McIntosh, Emma and Jordan Louviere (2002). Separating weight and scale value: an exploration of best-attribute scaling in health economics. Paper presented at Health Economics Study Group. Odense, Denmark.

Myers, James H. (1996) *Segmentation and positioning for strategic marketing decisions*. Chicago: American Marketing Association.

Paulhus, D.L. (1991). Measurement and control of response bias. J.P. Robinson, P.R. Shaver, and L.S. Wright (eds.), *Measures of personality and social psychological attitudes,* Academic Press, San Diego, CA.

Punj, Girish N. and David W. Stewart (1983). “Cluster Analysis in Marketing Research: Review and Suggestions for Application.” Journal of Marketing Research, 20, 134-48.

SAS Institute (2002). *The SAS System for Windows, SAS Institute*, Cary, North Carolina.

Steenkamp, Jan-Benedict E.M. and Frenkel Ter Hofstede (2002). International Market Segmentation: Issues and Outlook. *International Journal of Research in Marketing*, 19, 185-213.

Steenkamp, Jan-Benedict E.M. and Hans Baumgartner (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research* 25, 78-90.

Ter Hofstede, Frenkel, Jan-Benedict E.M. Steenkamp, and Michel Wedel (1999), International Market Segmentation Based on Consumer-Product Relations. *Journal of Marketing Research*, 36, 1-17.

Wedel, Michel and Wagner Kamakura. (1999). *Market Segmentation: Conceptual and Methodological Foundations*. Dordrecht: Kluwer Academic Publishers.

Wedel, Michel and Wayne S. DeSarbo. (2002) Market segment derivation and profiling via a finite mixture model framework. *Marketing Letters*, 13 (1), 17-25.