Sawtooth Software

RESEARCH PAPER SERIES

Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation

Steven H. Cohen,
SHC & Associates

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

Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation

Steven H. Cohen
SHC & Associates

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 versus 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. Product benefits are measured and then people with similar sets of benefits are termed "benefit segments." 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 three methods of measuring preferences for benefits using a split-sample design. Twenty benefits were presented to IT managers in an online survey. The first method uses a traditional ratings task. Each person performed 20 "mouse clicks" to rate the items on a 1-9 scale to fulfill the task. The second method uses 30 paired comparisons (cyclical design, chosen from the 20*19 = 380 possible pairs), yielding 30 mouse clicks. The third method uses Maximum Difference Scaling (described below), showing 20 sets of four benefits (quads) and asking the respondent to choose the Most Important and Least Important from each quad, resulting in 30 mouse clicks.

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*. We then present the results of the split-sample study described above. After that we describe how Maximum Difference Scaling can be combined with Latent Class Analysis to obtain benefit segments. We then describe an example of how both the traditional and the newer methods were used in a cross-national segmentation study of buyers of an industrial product conducted several years ago.

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.

Many of us have used rating scale data in factoring and in segmentation studies. The major problem tends to be response scale usage. Quite often we choose positively-worded important

items to include in a survey. The result is that the range of mean item scores is small and tends to be (at least in the USA) towards the top-end of the scale.

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 may get 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 becomes problematic when ratings have systematic scale use biases and large item inter-correlations owing to scale use. 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:

> *"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.

Even more egregious is the sensitivity of Cluster Analysis to the *order of the data*. Simply put, sort the data in one direction and obtain a solution. Then sort the data in the opposite way, specify the same number of clusters as in the first analysis. Now compare them. Our experience shows that using the clustering routines found in SAS and SPSS often yield an overlap of the two solutions in the 60% - 80% range. Not a very satisfying result, we contend.

Academic research has rightly pointed out other deficiencies of the two-stage or tandem approach of Factor Analysis followed by Cluster Analysis [see DeSarbo et al (1990); Dillon, Mulani, & Frederick (1989); Green & Krieger (1995); Wedel & Kamakura (1999); and, Arabie & 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 & 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 and attempt to dismiss Arabia & Hubert's concerns, the fact remains that their solution requires a heavy dose of analysis even before attempting to factor or cluster. The final segmentation analysis 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 section describes the use of Maximum Difference Scaling instead of rating scales to measure the *relative importance* of benefits and then we discuss the results of the split-sample study, comparing the IT managers' responses across ratings, simple paired comparisons, and the MaxDiff method.

We follow that section with a brief discussion of the advantages of Latent Class Analysis (LCA) over Cluster Analysis, as a method for uncovering market segments with similar benefit importances. We conclude with an illustration of using benefit segmentation with LCA in an international segmentation study of IT managers (different sample than the one used in the earlier analysis).

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 & 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 & 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.

- Using four items in the task (for example) and collecting the most and least in each task will result in recovering 5 of the 6 paired comparisons. For example, with items A, B, C, and D in a quad there are 4*3/2 = 6 pairs. If A were chosen as most and D as least, the only pair that we do not obtain a comparison of is the B-C pair.

- 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. Using HB methods will result in similar results as in an aggregate MNL model.

- Analyze pre-existing subgroups with the same statistical technique.

- To find benefit segments, use a Latent Class multinomial logit model.

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. Thus, one may think of MaxDiff as a more efficient way of collecting paired comparison data.

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.

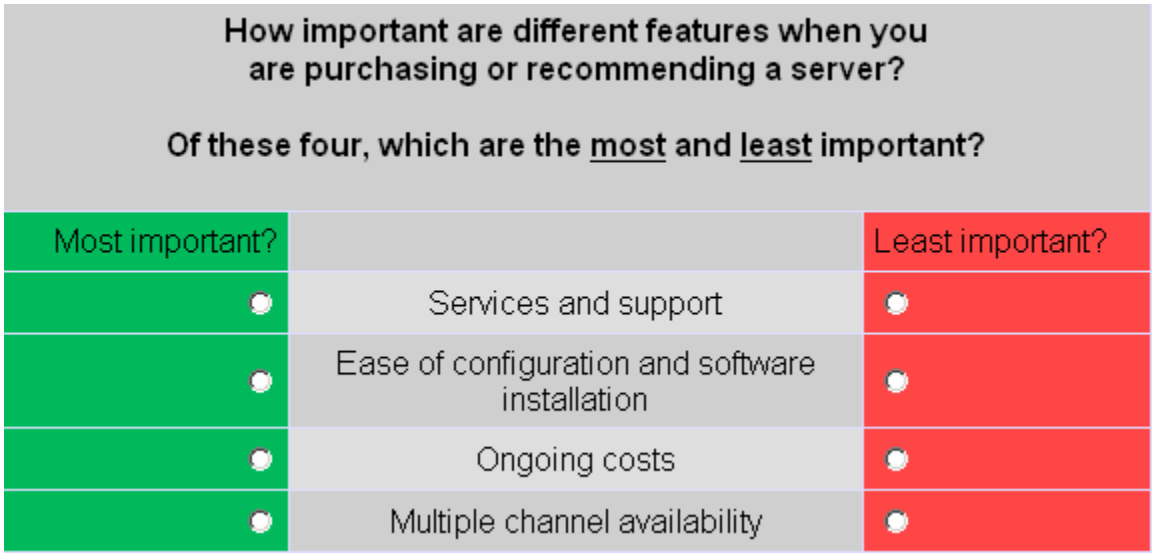
Each respondent will see each item in the first, second, third, etc. position across benefit subsets. The design will also control for context effects: each item will be seen with every other item an equal number of times.

The MaxDiff procedure will produce a unidimensional interval-level scale of benefit importance based on nominal level choice data. 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.

Comparisons of Results from the Three Methods Used in this Study

IT managers from an online panel were recruited and assigned to do one of the benefits evaluation tasks: 137 did ratings, 121 did paired comparisons, and 116 did the MaxDiff method.

Below is an example of a MaxDiff task for this study:



Immediately after the benefits evaluation, we asked the respondents to tell us their perceptions of the task they performed. As can be seen in Table 1, on a seven point scale of agree-disagree, all tasks were evaluated at about the midpoint of each scale, with ratings being slightly higher rated (e.g. easier) than the paired comparison or MaxDiff tasks. On average, the paired comparisons and MaxDiff task took about three times as long to complete than the ratings, but on the basis of "seconds per click," the ratings task took about ½ as long as the other two tasks, indicating the greater involvement and thought that is required.

Table 1
Qualitative Evaluation

Using a scale where a 1 means "strongly disagree" and a 7 means "strongly agree", how much do you agree or disagree that the <previous section>…

	Monadic (n = 137)	**Paired Comparison (n = 121)**	**Best/Worst (n = 116)**
…was enjoyable	4.3 (b, c)	4.0 (a)	3.8 (a)
…was confusing	2.4 (b, c)	2.9 (a)	3.2 (a)
…was easy	5.6 (b, c)	5.2 (a)	5.1 (a)
…made me feel like clicking answers just to get done	3.2	3.1 (c)	3.6 (b)
…allowed me to express my opinions	4.9 (c)	4.6	4.3 (a)

("a" means, significantly different from column a, $p<0.05$, etc.)

	Monadic (n = 137)	**Paired Comparison (n = 121)**	**Best/Worst (n = 116)**
Mean time to complete exercise	97 seconds	320 seconds	298 seconds
Seconds per mouse click	4.9 sec./click	10.7 sec./click	9.9 sec./click

The ratings task resulted in a 1-9 score for each of the 20 benefits. For each respondent, we chose 30 pairs (from three versions of a cyclic design) of the total number of 380 to be rated. Since MaxDiff requires two judgments per task, we chose 15 quads (from three versions of a computer-generated, balanced plan) for use in the MaxDiff task. Hence, we tried as best as we could to equalize the total number of clicks in the pairs and MaxDiff tasks. Both the paired comparison task and the MaxDiff task were analyzed using HB methods, resulting in 20 utilities for each person, typically ranging from about -4 to +4. The ratings task data suggested respondents often used a limited part of the scale. While the mean scores are very highly correlated across these two methods, the forced discrimination of the MaxDiff task should result in greater differentiation across items.

To test this hypothesis, we performed t-tests of mean benefit differences within each method. That is, we selected a reference item and compared each item's score to the reference item's score. We averaged the t-values obtained as a way to compare results, and we found that the average t-test for the rating scales was 3.3, for the paired comparison the average t-test result was 6.3, and the average for MaxDiff was 7.7. We conclude that the rating scale discriminated least among the items when comparing each one to the other, the MaxDiff results were most discriminating, and the paired comparison task was in between the other two, but closer to MaxDiff.

We then looked at the ability of each method to discover differences across pre-existing groups. We used 19 items from the survey, each with two to five categories and performed F-tests of

mean differences across the 19 items. Once again, we had 19*20 = 380 tests within method. By chance, we would expect that 19% of the tests (95% significance level) would be significant. Using the raw ratings data, we found 30 significant differences. Transforming the data to a within-person standardization (an often-used method to remove response biases) only yielded 22 significant differences. The paired comparison method yielded 40 significant differences, while the MaxDiff method resulted in 37, both about twice what would be expected by chance. Once again, we conclude the rating scales are less discriminating than the other two methods, but this time the paired comparison method performed a little better than MaxDiff.

We also gave each person four sets of three of the items as a holdout task, both prior to the scaling task and after (to assess test-retest reliability). We asked the person to rank-order the three items within each of the four sets. We then used the raw ratings or the utilities at the individual level to predict the rankings. Once again, MaxDiff was the winning method. As a percent of test-retest reliability, the hit rates were 97%, 88% and 85% for MaxDiff, Paired Comparisons, and Ratings respectively. While the performance of paired comparisons and ratings is commendable, the MaxDiff performance is quite astonishing, performing at about the same level as test-retest reliability.

We conclude that MaxDiff is certainly a superior method of collecting preferences than a ratings tasks. If we compare MaxDiff to paired comparisons, the evidence is that MaxDiff is superior, but not dramatically so.

Latent Class Analysis

We advocate using 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 LCA. In this case, the outcomes or response variables are the Most and Least choices from each set and the predictors are the presence or absence of each of the items in the set, and whether the item was chosen as most (coded +1) or chosen least (coded -1). 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 and from Statistical Innovations.

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

Our client, a multinational company offering industrial products around the globe, wished to conduct a study of its global customers. The goal of the research was to identify key leverage points for new product design and marketing messaging. Previous segmentation studies had failed to find well-differentiated segments and thus the marketing managers and the researchers were amenable to the use of the techniques described above. For the sake of disguising the product category and the client, we present the category as file servers.

The survey was administered in the client's three largest markets: North America, Germany, and Japan. 843 decision-makers were recruited for an in-person interview: 336 in North America, 335 in Germany, and 172 in Japan. The questionnaire contained background information on the respondent's company, their installed base of brands and products, and a trade-off task that examined new products, features, and prices. The benefit segmentation task is described next.

A list of thirteen product benefits was identified that covered a range of needs from product reliability to service and support to price. Prior qualitative research had identified these attributes as particularly desirable to server purchasers. The benefits tested were:

1. Brand name/vendor reputation
2. Product footprint
3. Expandability
4. Ease of maintenance & repair
5. Overall performance
6. Lowest purchase price
7. Redundant design
8. Reliability
9. Security features
10. Management tools
11. Technical support
12. Upgradeability
13. Warranty policy

A glossary was included with the survey so that respondents understood the meaning of each of these.

To develop the MaxDiff task, 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 item that was the most important and the least important when deciding which server to buy.

In this study, the utilities for the benefits range from positive 3.5 to negative 3.5. 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 exponentiation 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 in this study, 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 2, after rescaling.

Table 2
Overall Product Benefit Importances
from MaxDiff Task

Reliability	571
Overall Performance	277
Ease of Maintenance & Repair	84
Tech support	80
Expandability	59
Management tools	54
Upgradeability	50
Warranty policy	33
Brand name/reputation	27
Redundant design	24
Security features	27
Lowest Purchase Price	10
Product footprint	3

It is obvious that Product Reliability is the most important benefit followed by Overall Performance. In this market, Lowest Purchase Price and Product Footprint are the least important items. We then conducted a segmentation analysis of the Maximum Difference data using the Latent Class Multinomial logit model. A six segment solution was selected with the following segments emerging.

Table 3

Overall Product Benefit Importances
from MaxDiff Task by Benefit Segment

	Easy to Buy & Maintain	Never Breaks	Grows with Me	Help Me Fix It	Brand's the Clue	Managed & Safe
Reliability	264	601	373	554	623	481
Overall Performance	185	197	309	120	228	266
Ease of Maintenance & Repair	100	33	71	157	23	51
Technical support	86	34	34	305	23	58
Expandability	81	30	192	33	21	30
Management tools	53	29	38	23	26	190
Upgradeability	58	12	225	16	10	31
Warranty policy	100	14	21	45	20	29
Brand name/reputation	45	28	10	20	300	7
Redundant design	56	306	11	16	8	10
Security features	31	12	10	6	10	139
Lowest Purchase Price	213	3	5	4	7	5
Product footprint	28	1	2	1	2	3
Percent of total sample	17%	11%	19%	14%	16%	24%
Percent of expected product purchases	31%	19%	23%	9%	9%	9%

In all segments, Reliability is the most important benefit, but its importance varies greatly from a low index number of 264 in the first segment to a high of 623 in the fifth. The second most important benefit is Overall Performance, again ranging widely from 122 to 309. We would call these two "price of entry benefits" in the server category. Respondents in all segments agree, in varying intensities, that Reliability and Performance are what a server is all about. Segment differences reveal themselves in the remaining benefits.

- **Segment 1**, *Easy to Buy & Maintain* (17% of sample and 31% of future purchases), values Lowest Purchase Price (213), Ease of Maintenance & Repair (100), and Warranty Policy (100).

- **Segment 2**, *Never Breaks* (11% of sample and 19% of future purchases) values Redundant Design (306) even more than Performance (197). They have a high need for uptime.

- **Segment 3**, *Grows with Me* (19% and 23%), Values Upgradeability (225) and Expandability (192). They want to leverage their initial investment over time.

- **Segment 4**, *Help Me Fix It* (14% and 9%), values Technical Support (305) and Ease of Maintenance & Repair (157) even more than Performance.

- **Segment 5**, *Brand's the Clue* (16% and 9%), uses the Brand Name/Reputation (300) to help purchase highly reliable (623) servers. As the old saying goes, "No one ever got fired for buying IBM."

- **Segment 6**, *Managed & Safe* (24% and 9%), looks for Management Tools (190) and Security Features (139) when purchasing servers.

Note that Lowest Price, the second lowest index number overall is very important to the first segment, with an index score of 213. The benefits have very large variations across segments, indicating good between-segment differentiation. By looking at the number of servers expected to be purchased, we also provided guidance to management on which segments to target.

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 of rating scales and paired comparisons 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 the 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.

We conclude by quoting from Kamakura and Wedel's excellent book (1999) on Market Segmentation:

> *"The identification of market segments is highly dependent on the variables and methods used to define them."*

We hope that this paper demonstrates that the use of different scaling methods can influence the results of preference scaling and also segmentation research.

References

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

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

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

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. 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.

Cohen, Steven H. and Leopoldo Neira. (2003) "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.

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, & 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 partworths.” 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,” in 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.

Statistical Innovations, (2003). *Latent Gold. Statistical Innovations: Belmont, MA.*

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.
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.

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