A Procedure for Classifying New Respondents into Existing Segments Using Maximum Difference Scaling

Bryan Orme and Rich Johnson, Sawtooth Software
March, 2009 (with minor clarifications September 2016)

Background

Market segmentation is pervasive in our industry. Researchers are commonly called upon to develop needs-based or attitudinal segmentations based on respondent answers to questionnaires. The typical tools of the trade are Likert or semantic differential scales, followed by development of segments of like-minded respondents via cluster analysis, latent class analysis, or tree-based methods. There are many variations on these tools and the practice involves some art in addition to the science. Success is often measured by whether the segments have face validity, are stable (reproducible), have adequate size, provide insights for strategy, and are reachable, with meaningful differences on variables such as demographics and brand usage.

When managers incorporate a segmentation solution into their strategic thinking, they naturally want to profile respondents to *new* surveys into the same segments. Thus, segmentations are commonly accompanied by a *typing tool*: an abbreviated set of questions for classifying new respondents into existing segments. The methods commonly used to develop typing tools are discriminant analysis and tree-based algorithms. These identify the variables that best identify membership in segments and the mathematical or logical rules for assigning new respondents to segments.

Benefits of Maximum Difference Scaling (MaxDiff)

Respondents tend to use rating scales differently. Some tend to use just the high or the low scale points, and some conscientiously use the full breadth of the scale. If corrective measures aren't taken, the definition of segments may be strongly influenced by scale use bias rather than fundamental differences in preference. Another problem is that rating scales tend to be blunt and unreliable instruments. Ties are often registered, and many respondents cannot reliably map their internal feelings to what may seem to them an unnatural scale. Best-worst (Maximum Difference Scaling) is an excellent new approach that helps overcome these difficulties (Louviere 1991; Finn and Louviere, 1992).

Figure 1: Typical MaxDiff Question

When considering eating at a fast food restaurant, among the four attributes shown here, which of these is the <u>most</u> and <u>least</u> important?

Most Important		Least Important
○	Reasonable prices	○
○	Healthy food choices	○
○	Has a play area	○
○	Clean bathrooms	○

With Maximum Difference (MaxDiff) questionnaires, respondents are shown sets of typically four or five items and are asked to pick which item is best and which is worst within each set (see Figure 1). Typically, enough sets are shown to each respondent so that each item is seen multiple times (e.g. three or more times per respondent, if robust individual-level scores are desired). The more times an item is selected best (and not worst), the higher the resulting score. The scores are typically developed using logit, latent class, or hierarchical Bayes (HB) analysis—though simple counting approaches also can work nearly as well. Commercial software is available to perform the analysis, along with some free software (for advanced analysts with programming skills) such as the R statistical language.

Researchers have found that MaxDiff provides excellent data for conducting segmentations since it is free from scale-use bias and the resulting scores show strong discrimination on the items and larger differences between respondents than typical rating scales (Cohen and Orme, 2004). Despite the many advantages of MaxDiff questionnaires for preference measurement and segmentation, many researchers shy away from it because of the difficulty of developing a MaxDiff typing tool, should the client require one. The problem is more complex than developing a typing tool from rating scale questionnaires (where discriminant analysis and tree-based methods tend to work well to identify the key discriminating items). Not only do the items that most discriminate among segments need to be identified, but one has to decide how to arrange these items into efficient MaxDiff sets (consisting typically of four or five items per set). Finally, one needs to assign new respondents into existing segments based on choices of bests and worsts, and the math (though straightforward) is not as widely known for doing that.

Generating Efficient MaxDiff Sets for Classifying Respondents

Developing a MaxDiff typing tool involves a design optimization problem, where the goal is to find the most efficient set of MaxDiff questions for assigning respondents into an existing segmentation scheme. These questions may involve combinations of items from the original questionnaire, but probably new combinations not seen by the original respondents.

We have employed a relatively simple way to gauge how efficiently any potential set of MaxDiff questions can assign respondents into existing segments. Our approach requires individual-level scores on all the items from the original, full MaxDiff questionnaire. These must be logit-scaled (such that the antilog of the scores is proportional to choice likelihoods), and can be developed using a method such as hierarchical Bayes (HB). (HB uses choices from each respondent together with population means and covariances of the scores to obtain robust score estimates for each individual.) With those individual-level scores in hand, we can simulate how any respondent in the original dataset would answer a new and abbreviated MaxDiff typing questionnaire. The logit rule (and Bayesian logic) provides a way to determine to which segment a respondent most likely belongs, given the pattern of choices of best and worst that the respondent would be projected to make. If our typing tool assignment matches the assignment based on the original questionnaire, we count this as a correct classification *hit*. The goal is to design a typing questionnaire that will (in just a few questions) result in the highest hit rate for classifying respondents into known, original segments. Not surprisingly, the larger the typing questionnaire (more sets and more items per set), the more accurate the assignment. But, we find that the gains in classification accuracy diminish rapidly after just a few sets.

Finding an optimal typing questionnaire involves dealing with a truly immense search space. Imagine we wish to develop a typing questionnaire with just four MaxDiff sets, each involving five items. If the original questionnaire involves 25 items and we don't allow an item to be repeated within a set, there are 53,130 possible ways to design *just the first set*. There are 1.4 billion ways to design just the first *two* sets. For designing four sets, there are far too many combinations to examine exhaustively, one-by-one, and compute hit rates. So, we employ a fast swapping procedure that, although not guaranteed to find the globally optimal typing questionnaire, will identify near-optimal ones. It does so by starting with randomly designed sets of items, and for each set examining whether swapping each included item with each excluded item (considered one-at-a-time) would result in an improvement in hit rate. This swapping procedure continues until no further swaps can result in an improvement in hit rate. It typically solves the problem in a matter of seconds, and it should be repeated from multiple starting points to ensure a near-optimal solution.

Over repeated runs with the swapping algorithm, one can investigate the expected effectiveness (classification hit rate, based on the original respondents, using training and holdout samples) of different typing questionnaires that include differing numbers of sets and numbers of items per set. Although it isn't a requirement, typically one uses a number of items per set quite similar to the original questionnaire.

Assigning Respondents to Segments

We've avoided a detail in our previous explanation of the swapping algorithm that we now should explain: the method of assigning new respondents to segments based on answers to MaxDiff questions.

Imagine we have established a three-group segmentation scheme, and we've developed a short MaxDiff questionnaire for assigning new respondents into one of the three segments. The Latent Class theory provides the Bayesian framework for doing so.

Suppose the first set in a MaxDiff questionnaire includes four items: A, B, C, and D, and a new respondent chooses item A as best and item B as worst. The problem, then, is to determine to which of the three segments this respondent would most likely belong.

Assume the average logit-scaled scores[1] for segment 1 for items A, B, C, and D are:

A	1.0
B	-0.5
C	1.2
D	0.2

The logit rule specifies how to compute the likelihood that item A would be selected as best by members of this segment. We take the antilog (exponentiate) each of the item scores and normalize the scores as probabilities summing to 1.0.

Table 1: Likelihood that Each Item Selected "Best" for Segment 1

	Raw Scores	Exponentiated Scores	Probabilities
A	1.0	2.72	0.35
B	-0.5	0.61	0.08
C	1.2	3.32	0.42
D	0.2	1.22	0.15
		7.87	1.00

The likelihood that respondents belonging to segment 1 would choose item A as best is 0.35, or 35%.

Similarly, we can estimate the likelihood of respondents in segment 1 choosing B as worst from among the set of remaining items (B, C, and D). We do this by multiplying the raw scores by -1 (since identifying the item that is "worst" is polar opposite from identifying the "best"), and again following the logit rule.

---

[1] In early work with this typing tool search algorithm we employed average HB utilities for the segment for this purpose. Soon thereafter, we discovered that using aggregate logit scores for the respondents within the segment would obtain slightly better classification results.

Table 2: Likelihood that Each Item Selected "Worst" for Segment 1 (If A is Chosen "Best")

	Raw Scores	Exponentiated Scores	Probabilities
B	0.5	1.65	0.59
C	-1.2	0.30	0.11
D	-0.2	0.82	0.30
		2.77	1.00

The likelihood that respondents from segment 1 would choose B as worst within that set of three items is 0.59.

The likelihood of two independent events occurring is the product of their likelihoods. Of course, selections from a MaxDiff questionnaire are not truly independent, but it is common for market researchers to assume so. Thus, the likelihood of respondents belonging to segment 1 making that pattern of choices (A is best, B is worst) is (0.35)(0.59) = 0.21.

The size of segment 1 serves as a prior likelihood for new respondents (drawn from the same sample frame) belonging to segment 1. If segment 1 represents 40% of the population, then the posterior likelihood that a respondent picking A as best and B as worst belongs to segment 1 is proportional to (0.40)(0.35)(0.59) = 0.08. If more than one MaxDiff set has been answered, we continue to multiply the probabilities across MaxDiff sets.

We follow the same pattern of computation to predict the (relative) likelihood that this respondent belongs to segments 2 and 3. But, of course, the average logit scores (from an aggregate logit solution computed for the respondents belonging to the segment) for respondents belonging in segments 2 and 3 (along with the relative sizes of these segments) are substituted into the procedure above. We end up with three probabilities, each representing the relative posterior likelihood that a new respondent making this pattern of choices to the MaxDiff typing questionnaire would belong to each of the existing segments. We assign the new respondent to the segment reflecting the highest probability of membership.

Although we previously described the search routine as one that maximizes hit rates, simple hit rate is a step-shaped function and a less effective criterion to use in our search procedure (step-shaped functions are less informative regarding whether a proposed move in a given direction provides improvement or not; they often return a flat indication of progress until a certain threshold is reached). Rather, we rely on the likelihood that respondents are assigned into their original groups, which is a smooth function.

It is also possible to include demographic (or usage) information regarding respondents to assist with segment classification (often called covariates). The percentage of respondents of different demographic characteristics within each segment can be included in the likelihood computations in both the design and classification phases. If the covariates were not used as basis variables in developing the original segmentation, then this information typically provides only minimal lift in classification accuracy. But, if the covariates were used as basis variables in developing the original segmentation, adding that information to the typing tool can substantially improve classification. All of these computations can be performed instantaneously, even for real-time assignment of respondents to segments within web-based interviewing platforms.

An Empirical Test

In Feburary 2009, we conducted an empirical study to see if the MaxDiff typing tool approach we developed would work well for assigning respondents into previously-established segments. Using the hotspex Internet Panel, we interviewed about 700 Canadian respondents regarding 30 items related to the economy, political issues, and their personal financial position. The MaxDiff questionnaire asked respondents to indicate which items were most and least important in boosting their feelings of consumer confidence. The MaxDiff questionnaire consisted of eighteen sets with five items presented per set.

We used the MaxDiff importance scores (estimated via HB, then normalized) to develop a 5-segment solution (using cluster ensemble analysis), with relative sizes: 14%, 27%, 36%, 16%, and 7%. Based on the importance scores and the segmentation solution, we used the method described above to develop an abbreviated MaxDiff typing questionnaire consisting of six sets with five items per set (a third as long as the original questionnaire). We investigated other potential typing questionnaire solutions that differed in numbers of sets and numbers of items per set, but chose the combination that offered near-optimal classification with significantly reduced respondent effort (compared to the original, full MaxDiff questionnaire).

Two days after the data were collected, we posted the new MaxDiff typing questionnaire and invited the same respondents who completed the first wave of the study to return to complete the typing questionnaire. After a few more days of fielding, we had 556 respondents who had completed both waves of the study.

The challenge for the typing tool was to see if it could reassign each respondent into his/her original group as assigned in the first wave of the study. Reclassifying respondents correctly is not trivial, since respondents answer MaxDiff questions with some random degree of error, and there is even the possibility that respondents' opinions may have shifted in the few days between questionnaire waves. Also, the reality is that segments aren't always cleanly delineated, and there are often substantial numbers of respondents in the "troughs" of the distribution, distant from segment concentrations in the "peaks."

The typing tool was able to classify 60% of the respondents into the same segment. This represents a classification rate 3x greater than chance (there is a 20% likelihood of assigning a respondent correctly into 5 groups by random assignment).

What was also gratifying about the results of the empirical study was that we had simulated (via a series of split-sample validation steps using only wave 1 data) that new respondents (answering with error consistent with the logit rule) would be assigned to their actual segments with a 57% hit rate. This nearly matched what we actually observed in our empirical experiment.

If the goal in using the typing tool is to screen new respondents and identify members of certain segments with high accuracy, there is yet another way to boost the classification rate. Along with a prediction into a group, the typing tool also reports the likelihood that the respondent belongs to that group. If we isolate the 231 respondents who have at least 90% predicted likelihood of belonging to their predicted segment, the actual hit rate increased from the base rate of 60% to 77%. Respondents with 95% likelihood (170 respondents) were classified with 81% accuracy.

Table 3: Classification Accuracy into 5-Group Solution

Predicted Likelihood	Chance/Random Assignment	Assignment Accuracy
95% or better	20%	81%
90% or better	20%	77%
80% or better	20%	72%
70% or better	20%	70%
All Respondents	20%	60%

Extending to Other Choice Data

Our methodology isn't limited just to MaxDiff data. The common discrete choice questionnaires (choice-based conjoint) could also employ this approach. The extension simply involves designing combinations of product concepts (each composed of *multiple* attributes) within the choice sets rather than combinations of single items. Even though the search space is much larger, the swapping procedure described earlier can make quick work of finding near-optimal typing questionnaires.

Summary

Maximum Difference Scaling (MaxDiff) is a powerful tool for attribute/item scaling and segmenting respondents. The choices of best and worst items are free from scale use bias and provide strong discrimination on the items. If MaxDiff sets are so valuable for scaling data and segmenting respondents, then it also follows that MaxDiff sets should be equally useful components of typing tools, for assigning new respondents to existing segments. We have described a simple, yet powerful way for developing MaxDiff typing questionnaires and assigning new respondents into existing segments. With just six MaxDiff sets, classification accuracy into a 5-group segmentation solution for our empirical study is 60% (3x the chance level). Accuracy of assignment is boosted to 81%

by isolating respondents who have at least a 95% predicted likelihood of belonging to their predicted segments.

We hope the ideas and tools presented here will encourage more researchers to apply MaxDiff to segmentation work, and provide a clear roadmap for developing MaxDiff-based typing tools to assign new respondents into existing segments. More details regarding the approach may be obtained by writing the authors.

References

Cohen, Steve and Bryan Orme (2004), “What's Your Preference?” *Marketing Research,* 16 (Summer 2004), 32-37.

Finn, A. and J. 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, 12-25.

Louviere, J. J. (1991), “Best-Worst Scaling: A Model for the Largest Difference Judgments,” Working Paper, University of Alberta.