Sawtooth Software

TECHNICAL PAPER SERIES

The MaxDiff System Technical Paper

Version 8

Sawtooth Software, Inc.

1457 E 840 N Orem, Utah
+1 801 477 4700
www.sawtoothsoftware.com

The MaxDiff System Technical Paper
Sawtooth Software, Inc.
February, 2013

Introduction:

The MaxDiff System is software for obtaining preference/importance scores for multiple items (brand preferences, brand images, product features, advertising claims, etc.) using marketing or social survey research. MaxDiff is a component within Sawtooth Software's SSI Web platform for Web- and CAPI-based interviewing. It may be used designing, fielding, and analyzing:

- Method of Paired Comparisons (MPC) experiments (choices from pairs)
- Choices from subsets of three items, four items, etc.
- MaxDiff (best-worst scaling) experiments

Projects may be conducted over the Internet, using computers not connected to the internet (CAPI interviewing), or via paper-and-pencil questionnaires.

The Method of Paired Comparisons (MPC) is a very old and well-established approach for eliciting tradeoffs among paired items, dating back at least to the early 1900s (Thurstone 1927), with one author (David 1969) quoting a source as early as the 1800s (Fechner 1860). One can extend the theory of choices from pairs to choices among larger sets (three items, four items, etc.). MaxDiff (Maximum Difference scaling) is a successful, relatively new method for scaling multiple items (Louviere 1991, Finn and Louviere 1992). It may be thought of as a more sophisticated extension of MPC.

Our MaxDiff software makes it easy for researchers with only minimal exposure to statistics or advanced statistical methods to conduct sophisticated research involving the scaling of multiple items. For completeness of exposition, this document assumes that the reader has some background in statistics and multivariate methods such as hierarchical Bayes and Latent Class. If you are not familiar with these topics, don't despair. The software manual is much less technical than our explanation here. The trade-off techniques used in the MaxDiff System are very robust and easy to apply (for example, much easier to use than the related conjoint analysis). The resulting item scores are also easy to interpret. You do not need to have any formal training in statistics to use our MaxDiff System well and to achieve good results.

System Specifications:

Number of Items (Lab System)	Up to 20
Number of Items (30 System)	Up to 30
Number of Items (500 System)	Up to 500
Number of Sets per Respondent	Up to 200
Number of Questionnaire Versions across Respondents	Up to 999
Data Collection Modes	Web, Paper, CAPI
Parameter Estimation	hierarchical Bayes (HB), Logit, Latent Class

Motivation for MaxDiff:

Researchers in many disciplines face the common task of measuring the preference or importance of multiple items, such as brands, product features, employee benefits, advertising claims, etc. The most common (and easiest) scaling approaches have used rating, ranking, or chip allocation (i.e. constant sum tasks). (This software does *not* employ any of these approaches.)

Example Rating Task:

Please rate the following in terms of importance to you when eating at a fast food restaurant. Use a 10-point scale, where "0" means "not important at all" and "10" means "extremely important"

___ Clean bathrooms
___ Healthy food choices
___ Good taste
___ Reasonable prices
___ Has a play area
___ Restaurant is socially responsible
___ Courteous staff

Example Ranking Task:

Please rank (from most important to least important) the following in terms of importance to you when eating at a fast food restaurant. Put a "1" next to the most important item, a "2" next to the next most important item, etc.

___ Clean bathrooms
___ Healthy food choices
___ Good taste
___ Reasonable prices
___ Has a play area
___ Restaurant is socially responsible
___ Courteous staff

Example Allocation Task:

Please tell us how important the following are to you when eating at a fast food restaurant. Show the importance by assigning points to each attribute. The more important the attribute, the more points you should give it. You can use up to 100 total points. Your answers must sum to 100.

___ Clean bathrooms
___ Healthy food choices
___ Good taste
___ Reasonable prices
___ Has a play area
___ Restaurant is socially responsible
___ Courteous staff

Total: _____

There are a variety of ways these questions may be asked, including the use of grid-style layout with radio buttons and drag-and-drop for ranking. And, it is commonplace to measure many more items than the seven shown here. (These examples are not intended to represent the best possible wording, but are meant to be representative of what is typically used in practice.)

The common approaches (rating, ranking, and allocation tasks) have weaknesses.

- Rating tasks assume that respondents can communicate their true affinity for an item using a numeric rating scale. Rating data often are negatively affected by lack of discrimination among items and scale use bias (the tendency for respondents to use the scale in different ways, such as mainly using the top or bottom of the scale, or tending to use more or fewer available scale points.)

- Ranking tasks become difficult to manage when there are more than about seven items, and the resulting data are on an ordinal scale only.

- Allocation tasks are also challenging for respondents when there are many items. Even with a manageable number of items, some respondents may have difficulty distributing values that sum to a particular value. The mechanical task of making the allocated points sum to a particular amount may interfere with respondents revealing their true preferences.

Researchers seek scaling approaches that feature:

- Ease of use for respondents possessing a variety of educational and cultural backgrounds
- Strong discrimination among the items
- Robust scaling properties (ratio-scaled results preferred)
- Reduction or elimination of scale use bias

A very old approach, the Method of Paired Comparisons (MPC) (David 1969), seems to perform well on all these requirements. With MPC, respondents are shown items, two at a time, and are asked which of these two they prefer (or which is most important, etc.).

When considering eating at a fast food restaurant, which of these attributes is most important?

Reasonable prices	Healthy food choices
○	○

The respondent is not permitted to state that all items are equally preferred or important. Each respondent is typically asked to evaluate multiple pairs, where the pairs are selected using an experimental plan, so that all items have been evaluated by each respondent across the pairs (though typically not all possible pairings will have been seen) and that each item appears about an equal number of times. A study by Cohen and Orme showed that MPC performed better than the standard rating scale in terms of discrimination

among items and predictive validity of holdout (ranking) questions (Cohen and Orme 2004).

It should be noted that MPC can be extended to choices from triples (three items at a time), quads (four items at a time) or choices from even larger sets. There would seem to be benefits from asking respondents to evaluate three or more items at a time (up to some reasonable set size). One research paper we are aware of (Rounds *et al.* 1978) suggests that asking respondents to complete sets of from 3 to 5 items produces similar results as pairs (in terms of parameter estimates), and respondents may prefer completing fewer sets with more items rather than more sets with just two items.

Cohen and Orme also showed that a much newer technique, MaxDiff, may perform even better than MPC, especially in terms of predictive accuracy (Cohen and Orme 2002). MaxDiff questionnaires ask respondents to indicate both the most and least preferred/important items within each set.

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

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

Interest in MaxDiff has increased recently and papers on MaxDiff have won "best presentation" awards at recent ESOMAR and Sawtooth Software conferences (Cohen and Markowitz 2002, Cohen 2003, and Chrzan 2004).

What Is MaxDiff?

MaxDiff is a technique invented by Jordan Louviere in 1987 while on the faculty at the University of Alberta (Louviere, Personal Correspondence, 2005). The first working papers and publications occurred in the early 1990s (Louviere 1991, Finn and Louviere 1992, Louviere 1993, Louviere, Swait, and Anderson 1995). With MaxDiff, respondents are shown a set (subset) of the possible items in the study and are asked to indicate (among this subset with a minimum of three items) the best and worst items (or most and least important, etc.). MaxDiff (along with "best" only choices from sets of items) represents an extension of Thurstone's Law of Comparitive Judgement (Thurstone 1927).

According to Louviere, MaxDiff assumes that respondents evaluate all possible pairs of items within the displayed subset and choose the pair that reflects the maximum difference in preference or importance (Louviere 1993). However, the theory we prefer to apply is that respondents scan the set for the highest and lowest preference items (the

best-worst model specification).

MaxDiff may be thought of as a more sophisticated extension of the Method of Paired Comparisons. Consider a set in which a respondent evaluates four items, A, B, C and D. If the respondent says that A is best and D is worst, these two "clicks" (responses) inform us on five of six possible implied paired comparisons:

A>B, A>C, A>D, B>D, C>D

where ">" means "is more important/preferred than."

The only paired comparison that we cannot infer is B vs. C. In a choice among five items, MaxDiff questioning informs on seven of ten implied paired comparisons.

MaxDiff questionnaires are relatively easy for most respondents to understand. Furthermore, humans are much better at judging items at extremes than in discriminating among items of middling importance or preference (Louviere 1993). And since the responses involve choices of items rather than expressing strength of preference, there is no opportunity for scale use bias (MaxDiff is "scale free") (Cohen and Markowitz 2002). This is an extremely valuable property for cross-cultural research studies.

Analyzing Choice Data

The goal in using MaxDiff or the Method of Paired Comparisons is to achieve importance or preference scores for each item. The higher the score, the more important or stronger the preference.

The recent development of latent class and especially hierarchical Bayes (HB) estimation make methods such as the Method of Paired Comparisons and MaxDiff all the more appealing. That is because these methods employ choice data, which are sparse. Choices are natural for respondents to provide and are scale free, but they contain significantly less information for estimating scores than rating scales. Choice data reveal which item is preferred (or rejected as "worst"), but don't convey the intensity of preference. For the first seventy years or so, paired comparisons data were usually analyzed in the aggregate. The availability of Latent Class and HB extends our analysis capability considerably.

Latent class and HB make it possible to estimate stable item scores from relatively sparse choice data. They do so by borrowing information across the entire sample to stabilize the scores for segments or individuals.

- Latent class applied to MaxDiff data is a powerful approach for finding segments of respondents with quite differing preferences/importance scores (Cohen 2003). Latent Class should be more successful than clustering using data from standard rating scales. The MaxDiff System includes a Latent Class option that may be used with MaxDiff data.

- With HB modeling, we can derive useful scores at the individual level even though we have asked each respondent to evaluate only a fraction of all

possible subsets of items. With individual-level scores, researchers may apply common tests and tools, such as t-tests, cross-tabulations, histograms, and measures of dispersion (i.e. standard deviation). Although it is possible to submit individual-level scores derived from HB to a cluster procedure to develop segments, directly using Latent Class to simultaneously estimate item scores and segment respondents will probably yield better results.

- The MaxDiff System also includes an Aggregate Logit option, in case the user wishes to pool all respondent data together to compute average scores across the population. This is useful for quick, topline results, or to analyze especially sparse MaxDiff data (e.g. many items and few exposures of each item to each respondent).

Designing Paired Comparison Experiments

Optimal experimental designs for paired comparisons feature:

- **Frequency balance.** Each item appears an equal number of times.
- **Orthogonality.** Each item is paired with each other item an equal number of times.
- **Connectivity.** A set of items is connected if the items cannot be divided into two groups where any item within one group is never paired with any item within the other group. As a simple illustration, imagine four items: A, B, C and D. Assume that we ask just two paired comparisons: AB and CD. We could place each pair in a separate group and we could not determine the relative preferences across the items, since, for example, A was never used within the pairs of the other group. In contrast, if we had asked pairs AB, BC, and CD, then all items would be interconnected. Even though many pairs (such as AC and AD) had not been asked, we could infer the relative order of preference of these items.
- **Positional balance.** Each item appears an equal number of times on the left as it does on the right.

The MaxDiff System uses a cyclical algorithm to generate near-optimal plans for paired comparison experiments. Importantly, respondents receive multiple versions (blocks) of the plan, which tends to lead to more precise estimates (relative to a single version) and can reduce context and order effects. Consider a typical study with 24 items. There are k(k-1)/2 possible paired comparisons, or (24)(23)/2 = 276 possible pairs. Fortunately, each respondent does not need to complete all possible paired comparisons to lead to stable estimates of item scores.

At a minimum, we'd suggest asking each respondent 1.5x as many paired comparisons as items in the experiment (each item is shown three times). You will probably want to increase this by some amount, and the MaxDiff System gives great flexibility for choosing the exact number of pairs that you feel fits well within your questionnaire length and limitations. But, how should we choose the subset of the total number of possible paired comparisons to ask a respondent?

The MaxDiff System uses a cyclical approach for choosing designs. To illustrate, consider a small eight-item problem. There are (8)(7)/2 = 28 possible paired comparisons. We can organize the possible paired comparisons into efficient, progressive partitions (groups of questions). The first partition is the most valuable series of questions, the second partition is next-useful, and so on, until all paired comparisons have been described. Within each partition, the questions feature frequency balance.

For the eight-item experiment, there are four possible frequency-balanced partitions:

Partition 1	Partition 2	Partition 3	Partition 4
1 2	1 5	1 4	1 3
2 3	2 6	2 5	2 4
3 4	3 7	3 6	3 5
4 5	4 8	4 7	4 6
5 6		5 8	5 7
6 7		6 1	6 8
7 8		7 2	7 1
8 1		8 3	8 2

Partition 1 completes a full circle, pairing each item with its adjacent item. Partition 1 achieves both frequency balance (each item shown twice) and full connectivity of the items. Partition 2 pairs each item with the item half-way down the list. It achieves perfect frequency balance with just four pairings (each item shown once). However, by itself, it doesn't achieve connectivity of the items (but we already established connectivity using Partition 1). Partitions 3 and 4 also reflect frequency balance, though only Partition 3 reflects connectivity. All together, there are 28 paired comparisons across the four partitions, reflecting the full-factorial (all possible pairs).

If enough questions are asked (at least as many questions as items in the study) the cyclical design approach ensures that each version of the design (respondent questionnaire) features connectivity. If you need to ask each respondent fewer questions than the number of items in the study, it becomes very challenging for HB to scale the items relative to one another at the individual level. With fewer questions per respondent than items in the study, we'd recommend using latent class modeling.

If we could only ask eight questions of a respondent, we should use Partition 1. If we could only ask twelve questions, we should ask Partitions 1 and 2. The next priority would be Partition 3 followed by Partition 4.

A similar pattern generalizes for all experiments involving an even number of items. (With ten items, there are five partitions in the full factorial, with twelve items there are six partitions, etc.) Partitions with odd integer differences between paired items reflect connectivity; those with even integer differences between paired items are not connected.

In the case of an odd number of items in the study, the same patterns are observed, except there is no partition similar to Partition 2 above, where a frequency-balanced block can be completed in 1/2k comparisons, where k is the number of items in the study (obviously, k is not evenly divisible by 2). Also, partitions with even integer differences between items

are connected; those with odd integer differences between paired items are not connected (with the exception of Partition 1).

The software is very flexible, allowing you to specify exactly how many questions to ask each respondent[*]. For example, if you specify that 18 questions should be asked for an 8-item experiment, Partitions 1 and 2, and six questions from Partition 3 would be chosen. After the paired comparisons have been selected for a questionnaire version, the question order is randomized.

Subsequent questionnaire versions are chosen such that the order of items taken into the partitions is randomized and (in the case of a question count that is not evenly filled using the partitions) deficits in item representation across the versions are attempted to be remedied in later versions. The algorithm also pays attention to two-way frequency of items (how many times each item appears with each other item). Subsequent questionnaire versions are chosen such that deficits in two-way frequencies are reduced in later versions. The algorithm also seeks to balance the number of times each item appears on the left and on the right.

Designing Experiments Involving Choices from Three or More Items

In MaxDiff questionnaires or in extensions to MPC where respondents are asked to choose just the most preferred/important item from sets of three or more items, we again must decide how to choose an efficient fraction of the total possible combinations of items to show each respondent.

The MaxDiff System uses a programming-based algorithm to choose designs according to the same criteria as mentioned previously with respect to MPC designs:

- **Frequency balance**
- **Orthogonality**
- **Connectivity**
- **Positional Balance**

As a final step, we randomize the order of the tasks within each version.

The design process is repeated 1000 separate times by default (using a different starting seed each time), and the replication that demonstrates the best one-way balance (number of times each item occurs) is selected. If multiple designs have the same degree of one-way balance, then we select among those designs based on the best two-way balance (number of times each pair of items occurs within sets). If multiple designs have the same degree of one-way and two-way balance, then we select among those designs based on the best positional balance. With small to moderate sized designs and no prohibitions, this usually happens within a few seconds.

The MaxDiff System produces a report to help you evaluate the quality of the design.

[*] This flexibility comes at a small penalty in certain rare cases as slightly more efficient MPC designs exist for numbers of pairs that are very specific multiples of the numbers of items (Grimshaw *et al*. 2001).

Here is an example involving just two questionnaire versions (such as might be implemented for a simple paper-and-pencil experiment, where it would be labor intensive to offer a unique version for each respondent):

```
Number of Items (Attributes): 8
Number of Items per Set: 4
Number of Sets per Respondent: 10
Number of Questionnaire Versions: 2
Random Number Seed: 1

One Way Frequencies:
1      10
2      10
3      10
4      10
5      10
6      10
7      10
8      10

Mean = 10.000000
Std Dev. = 0.000000

Two Way Frequencies:
       1      2      3      4      5       6      7      8

1      10     4      4      5      4      5      4      4
2      4      10     5      4      4      4      4      5
3      4      5      10     4      5      4      4      4
4      5      4      4      10     4      4      4      5
5      4      4      5      4      10     4      5      4
6      5      4      4      4      4      10     5      4
7      4      4      4      4      5      5      10     4
8      4      5      4      5      4      4      4      10

Off Diagonal Non-prohibited Elements
Mean = 4.285714
Std Dev. = 0.451754

Positional Frequencies:
       1      2      3      4

1      2      3      3      2
2      2      2      3      3
3      3      3      2      2
4      3      3      2      2
5      2      2      3      3
6      3      3      2      2
7      2      2      3      3
8      3      2      2      3

Mean = 2.500000
Std Dev. = 0.500000
```

In this example, the one-way frequencies are perfectly balanced and the off-diagonal two-way frequencies nearly so. It is not necessary to achieve *exact* balance in a design to have a very satisfactory design, but balance is a desirable property. Methods such as logit, latent class, and HB do not require perfect balance to achieve unbiased estimates of

parameters.

The Positional Frequencies report how many times each item appears in the first, second, third, and so on positions. A standard deviation is reported for each table. Lower standard deviations are better, with a standard deviation of zero reflecting perfect balance. With most design specifications (number of items, items per set, sets, and versions), it is impossible to achieve exact balance in all three tables. Fortunately, exact balance is not required for near-optimal efficiency and unbiased estimation of parameters. However, the researcher interested in achieving slightly better designs should try different starting seeds and perhaps even more than 1000 iterations, and compare the results.

How Many Items and Sets to Show

Research using synthetic data (Orme, 2005) suggests that asking respondents to evaluate more than about five items at a time within each set may not be very useful. The gains in precision of the estimates are minimal when using more than five items at a time per set for studies involving up to about 30 total items. Orme speculated that the small gains from showing even more items may be offset by respondent fatigue or confusion.

Another finding from Orme's research is that it is counterproductive to show more than half as many items within each set as are in the study. Doing so can actually decrease precision of the estimates. Orme provided this explanation: "To explain this result, consider a MaxDiff study of 10 items where we display all 10 items in each task. For each respondent, we'd certainly learn which item was best and which was worst, but we'd learn little else about the items of middle importance for each individual. Thus, increasing the number of items per set eventually results in lower precision for items of middle importance or preference. This leads to the suggestion that one include no more than about half as many items per task as being studied."

Orme's simulation study also included an internal validation measure using holdouts, leading to a suggestion regarding how many subsets to ask each respondent in MaxDiff studies. He stated, "The data also suggest that displaying each item three or more times per respondent works well for obtaining reasonably precise individual-level estimates with HB. Asking more tasks, such that the number of exposures per item is increased well beyond three, seems to offer significant benefit, provided respondents don't become fatigued and provide data of reduced quality."

Should We Ask for "Worsts"?

The MaxDiff System allows researchers to ask for "best" and "worst" choices within subsets of items (set size >=3), or to ask only for "bests." Collecting both bests and worsts contributes more information. However, it has been shown that the parameters resulting from best choices may differ (statistically significant differences) from those developed only using worst choices. Even so, the results tend to be quite similar between bests and worsts. There is some debate among leading academics regarding the statistical properties of "worsts" and whether including both bests and worsts is an appropriate extension of the logit model. We at Sawtooth Software do not know which approach is

best (to ask for "worsts" or not). We imagine that it depends on the focus of the research. For example, healthcare outcomes studies often are concerned more about identifying the worst outcomes with greater precision than the best outcomes. Asking only for "bests" is theoretically more sound, but asking for "worsts" seems to offer practical value. We hope that offering flexibility in this software will lead to more experimentation in this area.

If the purpose of the research is to measure the full range of scores from the best item to the worst, it seems to us useful to ask for both best and worst judgments. For example, if studying a variety of positions in a political campaign, it may as important to identify positive positions as to identify those that might have a negative impression for groups of potential voters. If the purpose of the research focuses on the trade-offs among items at the favorable end of the scale for each respondent (such as is the case in choice simulations or TURF analysis), then it may be more useful just to ask for bests and avoid spending time asking for worsts. The statements within this paragraph reflect our opinions at this point, and we have no data to support them.

Analyzing Results for MPC and Choice of "Best" Only Experiments

Counting Analysis

The simplest method of analysis consists of counting choices. For each respondent, choice set, and item, we code a score. If an item is available in the choice set and chosen, that item is coded as +1. If an item is available but not chosen, that item is coded 0. If the item is not available within the subset, it is coded as missing. In Counting analysis, we simply average these scores across all respondents and tasks, for each item. The resulting means are probabilities of choice, ranging from 0 to 1. The probability reflects the likelihood that the item is chosen within all possible subsets (of the size included in the questionnaire) including this item in the study. The probabilities do not add to 100%, since not all items are available within each choice set. The probabilities will sum to k/t, where k is the total number of items in the study and t is the number of items shown per set.

Counting analysis is a quick way to summarize the average preference or importance of items across respondents, under conditions of near-orthogonality. That is to say that each item should appear approximately an equal number of times with every other item in the experiment. With designs produced by the MaxDiff System, near-orthogonality holds when no prohibitions are specified between items and when many versions of the questionnaire are used. However, counts tend to be "flatter" than the more accurate assessment of scores given through Logit, Latent Class, or HB analysis.

Multinomial Logit and Hierarchical Bayes Modeling

For paired comparisons or choices (bests only) among sets with more than two items, we may treat the process as a utility-maximizing decision following Random Utility Theory, modeled using multinomial logit (MNL). We find a set of weights for the items such that when applying the logit rule we obtain a maximum likelihood fit to the respondents' actual choices. The logit rule specifies that the probability of choosing the *ith* item as

best (or most important) from a set containing *i* through *j* items is equal to:

$$P_i = e^{U_i} / \Sigma\, e^{U_{ij}}$$

where e^{U_i} means to take the antilog of the utility for item *i*.

To avoid linear dependency, we arbitrarily set the utility for the last item to zero and estimate the utility of all other k-1 items with respect to that final item held constant at zero. This is accomplished through dummy coding.

In the MaxDiff System, we employ hierarchical Bayes estimation to compute individual-level weights under the logit rule. Hierarchical Bayes borrows information across the sample to stabilize the estimates for each individual. We use a prior covariance matrix appropriate for proper estimation of categorical attributes to avoid potential troubles estimating the omitted level with respect to the other parameters, as described by Orme and Lenk (Orme and Lenk 2004). The interested reader can refer to the CBC/HB Technical Paper for more information on hierarchical Bayes modeling (Sawtooth Software 2004) and also Lenk's procedure for specifying a proper prior covariance matrix.

To make it easier for users to interpret the raw scores, we zero-center the parameters as a final step before saving them to file.

Probability-Based Rescaling Procedure

The weights (item scores) resulting from multinomial logit (MNL) estimation typically consist of both negative and positive values. These values are on an interval scale and their meaning is sometimes difficult for non-technical managers to grasp. Because these experiments use choice data, we can convert these scores to ratio-scaled probabilities that range from 0 to 100.

To convert the raw weights to the 0-100 point scale, we perform the following transformation for each item score:

$$e^{Ui} / (e^{Ui} + a - 1)$$

Where:

U_i = zero-centered raw logit weight for item i
e^{Ui} is equivalent to taking the antilog of U_i. In Excel, use the formula =EXP(U_i)
a = Number of items shown per set

Then, for convenience, we rescale the transformed values to sum to 100.

The logic behind this transformation is as follows: We are interested in transforming raw scores (developed under the logit rule) to probabilities true to the original data generation process (the counts). If respondents saw 4 items at a time in each MaxDiff set, then the

raw logit weights are developed consistent with the logit rule and the data generation process. Stated another way, the scaling of the weights will be consistent within the context (and assumed error level) of choices from quads. Therefore, if an item has a raw weight of 2.0, then we expect that the likelihood of it being picked within the context of a representative choice set involving 4 items is (according to the logit rule):

$$e^{2.0} / (e^{2.0} + e^{0} + e^{0} + e^{0})$$

Since we are using zero-centered raw utilities, the expected utility for the competing three items within the set would each be 0. Since $e^{0} = 1$, the appropriate constant to add to the denominator of the rescaling equation above is the number of alternatives minus 1.

The rescaled scores follow ratio scaling[†], and we may say that an item with a score of 10 is twice as preferred or important as an item with a score of 5. The results should be easily understood by non-technical people, as most people are used to seeing survey results where the results for items range from 0 to 100.

Analyzing Results for MaxDiff Experiments

Counting Analysis

As with paired comparisons, we can compute the probability that an item is chosen best or worst when it was available within a subset (see previous explanation). The MaxDiff System reports these probabilities separately, for bests and worsts.

Multinomial Logit and Hierarchical Bayes Modeling

For MaxDiff experiments, assume that choices of bests and choices of worsts may be treated as a utility-maximizing (minimizing) decision following Random Utility Theory, modeled using multinomial logit (MNL). We find a set of weights for the items such that when applying the logit rule we obtain a maximum likelihood fit to the respondents' actual choices.

Under the logit rule, the probability of choosing the *ith* item as best (or most important) from a set containing *i* through *j* items is equal to:

$$P_i = e^{U_i} / \Sigma\, e^{U_{ij}}$$

[†]This rescaling procedure does not factor out the scale parameter. Scale factor is inversely related to noise. The greater the noise, the "flatter" the probability scores. Thus, it is inappropriate to make direct comparisons between groups of respondents who differ significantly on scale. But, such is the case with many types of marketing research data, where the strength of signal depends upon the degree to which respondents pay attention or understand the question. Most every informed researcher willingly accepts this and expects noise to be fairly constant across respondent groups, such that this will make little matter when comparing groups. If this issue still concerns you, each respondent's scores might be re-scaled such that the mean is zero and the scores have a range of 100 points. This transformation comes at a cost, however. After rescaling, the scores will not retain their ratio-scaled properties and will no longer correspond to choice probabilities.

where e^{Ui} means to take the antilog of the utility for item i.

and the probability of choosing the *ith* item as worst (or least important) is equal to:

$$P_i = e^{-U_i} / \Sigma e^{-U_{ij}}$$

where e^{-U_i} means to take the antilog of *negative* the utility for item i.

Researchers have found that the parameters and scale factor from best and worst choices may vary (though the scale factor is more likely to vary than the relative preference weights). However, it is common to concatenate both kinds of data and estimate using a single MNL model, as described by Louviere (Louviere 1993), and that is the approach we take with the MaxDiff System. The independent variable matrix is dummy-coded with k-1 parameters (the last item is omitted and constrained to have a weight of zero). Each set is coded as two separate sets: one for best responses and one for worsts. All elements in the design matrix are multiplied by -1 in the set describing worsts. (Again, there is debate regarding the correctness of modeling MaxDiff in the way as described by Louviere, but the results seem to work very well in practice.)

We employ hierarchical Bayes to estimate individual-level weights under the logit rule, as described previously for paired comparison experiments. To make it easier for users to interpret the raw scores, we zero-center the parameters as a final step before saving them to file.

Probability-Based Rescaling Procedure

As previously described, the weights (item scores) resulting from multinomial logit typically consist of both negative and positive values. These values are on an interval scale and are sometimes difficult for non-technical individuals to grasp. Because MaxDiff experiments use choice data, we can convert these scores to ratio-scaled probabilities that range from 0 to 100 as described previously.

Respondents indicate which items are relatively better (or worse) than others within standard MaxDiff questionnaires. Thus, the scores are estimated on a relative scale, without any indication that the items are good or bad, important or unimportant, in an absolute sense. Some researchers have viewed that as a limitation for certain studies and certain clients.

Anchored MaxDiff

Anchored MaxDiff lets the researcher draw a line (a utility boundary) between important and unimportant items (positive vs. negative impact, buy or no-buy, etc.). That utility boundary, for example, could be set at 0. Thus, any items that are positive are considered important and those that are negative are not important (in an absolute sense). Anchored MaxDiff score estimation is available for aggregate logit, latent class, and HB score estimation routines within the MaxDiff System.

Anchored MaxDiff provides potentially more information than standard MaxDiff, but it

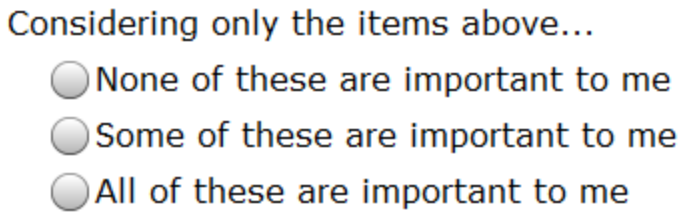
comes at a cost. One of the key benefits of standard MaxDiff is that it is free from scale use bias, making it a very nice technique for studying relative preferences across countries or across respondents who have different scale use bias tendencies.

With anchored MaxDiff, although a rating scale is not being used, the tendency for one group of respondents to generally be more positive/agreeable than another group of respondents can lead to similar problems as scale use bias, which was one of the main problems researchers have wanted to eliminate by using MaxDiff!

Furthermore, using anchored MaxDiff scores within cluster segmentation (or latent class segmentation) might lead to respondent groups that are being delineated as much by their tendency to be positive/agreeable regarding the position of items vs. the anchor as by their relative scores for the items of interest within your study. (An obvious solution to this second issue is to develop the segmentation using un-anchored scores, but then to profile the segments using the anchored scores.)

Dual-Response Indirect Method

Jordan Louviere, the inventor of MaxDiff scaling, proposed a dual-response, indirect method for scaling the items relative to a threshold anchor of importance or desirability. You may add indirect scaling questions to your MaxDiff survey by clicking the Add Dual-Response Question box from the Format tab within your MaxDiff exercise. Below each MaxDiff question, a second (hence, "dual") question is inserted, such as the following:



If the respondent clicks "None of these are important to me" then we inform utility estimation that all four items shown in this MaxDiff set should have lower utility than the anchor threshold.

If the respondent clicks "All of these are important to me" then we inform utility estimation that all four items should have higher utility than the anchor threshold.

If the respondent clicks "Some of these are important to me" then we know that the "best" item selected within the MaxDiff set should have higher utility than the anchor and the "worst" item selected should have lower utility than the anchor.

When "Some of these are important to me" is selected, we do not have any information about how the two non-selected items (from this set of four) relate to the anchor. Thus, the dual-response indirect method provides incomplete information about how each of the items shown in a MaxDiff set relates to the anchor.

Note: if the dual-response indirect method is applied to questions that show just two items per set (paired comparisons questions), then we achieve complete information regarding how the two items shown relate to the anchor.

The indirect method should probably not be used with more than 4 items shown per set. Increasing the number of items per set increases the likelihood that respondents will select "Some of these are important to me," which provides incomplete information. Because of this issue, the software issues a warning if you try to use the indirect method with more than 4 items displayed per MaxDiff set.

The indirect method works well in practice. Evidence has been presented at the Sawtooth Software Conference that the indirect method tends to lead to more items being scaled above the anchor threshold than the direct method. In other words, the indirect method leads to more items being judged as "Important" or a "Buy" compared to the direct method (described directly below).

Direct Binary Approach Method

Both Bryan Orme (in 2009 at the SKIM/Sawtooth Software European Conference) and Kevin Lattery (at the 2010 Sawtooth Software Conference) have demonstrated how standard ratings questions (or sorting tasks) may be used to anchor MaxDiff items. Lattery's work is more generally referenced today, as his paper is more complete in comparing direct and indirect anchoring methods. Using questions outside the MaxDiff section, we may ask respondents directly whether each item (or each of a subset of the items) is important or not. A 2-point scale could be used, or respondents could be asked to sort items into two buckets: important and unimportant buckets. A 5-point scale could also be used, where items rated either top box or top-two box could signal that these exceed the importance threshold boundary, etc.

The direct method works well in practice. Evidence has been presented at the Sawtooth Software Conference that the direct method tends to lead to more items being scaled below the anchor threshold than the indirect method. In other words, the direct method leads to more items being judged as "Not Important" or a "No buy" compared to the indirect method (described directly above). Context bias can become a potential problem with the direct method. If the number of items grows to the point that the researcher wishes to show the items across multiple screens, then the context of the other items on the screen can affect the absolute judgments given.

Validation Research-on-Research for Anchored Scaling MaxDiff

Sawtooth Software and Kevin Lattery (from Maritz) have separately done a few research-on-research projects to test and validate different methods for anchored scaling in MaxDiff. To read more about different approaches to direct and indirect anchored scaling approaches, please see:

- "Using Calibration Questions to Obtain Absolute Scaling in MaxDiff" (Orme 2009)

- "Anchored Scaling in MaxDiff Using Dual-Response" (Orme 2009)

- "Anchoring MaxDiff Scaling Against a Threshold - Dual Response and Direct Binary Responses" (Lattery 2010)

MaxDiff Scaling and "Best-Worst Conjoint" Analysis

Many of the first papers published on MaxDiff scaling suggested that it could be used with conjoint-style problems, where there are attributes with mutually exclusive multiple levels. Louviere and colleagues presented this idea under the title of "best-worst conjoint" (Louviere, Swait and Anderson 1995). Respondents were shown the product profile and asked which features made them most and least likely to want to buy the product. The authors argued that this approach could yield useful part worth parameters, where all the part worths were placed on a common scale (an advantage over standard conjoint). Unlike traditional conjoint analysis where it is not proper to compare the utility of a single level for Price with the utility for a single level of Speed, such comparisons could be made under best-worst conjoint.

So-called "best-worst conjoint" has not gained much traction in the industry. We at Sawtooth Software do not consider it a true conjoint method, as respondents are never asked to evaluate conjoined elements as a whole. Because respondents do not respond to the product concept as a whole, many researchers question whether part worth utilities from best-worst conjoint are appropriate to use in traditional conjoint simulators. However, in recent correspondence with us at Sawtooth Software, Louviere reports success in using the parameters from best-worst conjoint in market simulations (Louviere 2005). Most recently, we've become aware of additional research that lends credence to his claims. We've conducted our own methodological research and also seen reported by others that best-worst conjoint can perform almost as well as CBC in terms of predicting CBC-like holdouts. CBC is still superior (especially if interactions are of concern), but "best-worst conjoint" performs reasonably well.

It is possible, using a series of item prohibitions, to implement best-worst conjoint with the MaxDiff System. However, the software was not designed for this purpose.

Confidence Bounds and Statistical Testing

The MaxDiff System reports upper and lower 95% confidence bounds for the raw and rescaled scores. These are computed in the "classical" tradition by estimating the standard error for each item based on the respondents' point estimates, and adding +/- 1.96 standard errors to the mean.

We have chosen to report confidence bounds due to popular demand. Users have come to expect this and it is common practice in our industry. We recognize, however, that computing standard errors based on the point estimates is not true to the Bayesian tradition. Also, the rescaled scores reflect truncated, skewed distributions, for which developing confidence bounds is not as appropriate as when using the raw parameters (which tend to be more normally distributed).

If these issues concern you and you wish to conduct statistical tests and confidence bounds more to the Bayesian tradition, then we suggest you export the data for use with our CBC/HB product, which can produce a file of draws that may be used to develop more proper confidence bounds and statistical tests.

Error Theory and the MaxDiff Model

In this section, we use simulated data to demonstrate that MaxDiff data can conform to assumptions inherent in MNL and to Random Utility Theory.

Assume the respondent has some latent (unobservable) utility structure. To make the selection of "best," assume the respondent evaluates items in the set and chooses the highest utility item according to the utilities plus extreme value error (right-skewed Gumbel error). To choose the worst item, assume the respondent independently evaluates items in the set and makes a selection of "worst" according to the utilities plus extreme value error (*left*-skewed Gumbel error).

We have simulated this data generation process, and found that MNL (aggregate logit) can recover known utility parameters. We employed the following steps: We generated five thousand synthetic respondents with 12 parameters, assuming population means of -5.5, -5.0, -4.5, -4.0, -3.5, -3.0, -2.5, -2.0, -1.5, -1.0, -0.5, and 0. We generated a MaxDiff questionnaire with 12 sets, where each set contained 4 items. Our synthetic respondents answered the questionnaire according to the previously described rule. We coded the independent variable matrix as described earlier in this documentation.

Recovery of True Parameters in MaxDiff

True Mean	Estimated Mean
-5.50	-5.49
-5.00	-5.00
-4.50	-4.51
-4.00	-4.00
-3.50	-3.51
-3.00	-3.02
-2.50	-2.51
-2.00	-2.03
-1.50	-1.53
-1.00	-1.02
-0.50	-0.53

The estimated parameters are quite close to the parameters used to generate the data set. The small differences are explainable by random error.

It can be argued that the assumptions we made in generating the synthetic data really don't hold with actual MaxDiff questionnaires, since the choice of best and worst are not truly independent. They are exclusive for each task. But our synthetic respondents very occasionally chose the same items as best and worst for tasks (since independent Gumbel

error was used). If we change the decision rule for the synthetic respondents such that the choice of best and worst must be exclusive, and re-run the MNL estimation, the estimated utility values become a little larger in scale, reflecting essentially a uniform transformation of the previously estimated values when exclusivity of best and worst was not enforced. The interpretation of the parameters, for all practical purposes, would be the same.

We have demonstrated that an error theory consistent with MNL can be developed for MaxDiff. Humans do not behave as nicely as simulated respondents, and the previous cautions still apply regarding the blending of best and worst judgments.

References

Chrzan, Keith (2004), "The Options Pricing Model: A Pricing Application of Best-Worst Measurement," *2004 Sawtooth Software Conference Proceedings*, Sequim, WA.

Cohen, Steve 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, Steve (2003), "Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation," *2003 Sawtooth Software Conference Proceedings*, Sequim, WA.

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

David, H.A. (1969), The Method of Paired Comparisons, Charles Griffin & Company Ltd., London.

Fechner, G. T. (1860), *Elemente der Psychophysik*. Leipzig: Breitkopf and Hartel.

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.

Grimshaw, Scott D., Bruce J. Collings, Wayne A. Larsen, and Carolyn R. Hurt (2001), "Eliciting Factor Importance in a Designed Experiment," *Technometrics*, May 2001, Vol. 43, No. 2.

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

Louviere, J. J. (1993), "The Best-Worst or Maximum Difference Measurement Model: Applications to Behavioral Research in Marketing," The American Marketing Association's 1993 Behavioral Research Conference, Phoenix, Arizona.

Louviere, J. 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," Unpublished working paper, University of Sydney.

Louviere, Jordan (2005), Personal Correspondence with Bryan Orme via Email.

Orme, Bryan and Peter Lenk (2004), "HB Estimation for "Sparse" Data Sets: The Priors Can Matter" 2004 ART Forum, American Marketing Association, Whistler, BC.

Orme, Bryan (2005), "Accuracy of HB Estimation in MaxDiff Experiments," Technical Paper available at www.sawtoothsoftware.com.

Rounds, James B., Jr., Thomas W. Miller, and Rene V. Dawis, "Comparability of Multiple Rank Order and Paired Comparison Methods," Applied Psychological Measurement, Vol. 2, No. 3, Summer 1978, pp. 413-420.

Sawtooth Software (2004), "The CBC/HB Technical Paper," Technical Paper available at www.sawtoothsoftware.com.

Thurstone, L. L. (1927), "A Law of Comparative Judgment," Psychological Review, 4, 273-286.