Sawtooth Software

RESEARCH PAPER SERIES

Anchored Scaling in MaxDiff Using Dual Response

Bryan K. Orme,
Sawtooth Software, Inc.

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

Anchored Scaling in MaxDiff Using Dual-Response

Bryan Orme, Sawtooth Software

Despite the general enthusiasm regarding MaxDiff (Maximum Difference Scaling, Louviere 1991), some researchers have worried about the relative (ipsative) nature of the resulting scores. In a MaxDiff question, respondents are shown typically four or five items at a time, and they indicate for each set which item is best (or more important) and which is worst (or least important). But, respondents cannot say, for example, that *all* of the items are important or *all* of the items are not important. MaxDiff measures only *relative* desirability among the items.

Researchers have worried that relative preference scores omit meaningful information that distinguishes respondents or groups of respondents (Bacon *et al.* 2008). Whether this issue matters much in practice is the subject of some debate, and including a wide variety of items in a MaxDiff task reduces the potential problems of the relative scale.

Advanced statistical procedures have been proposed to anchor MaxDiff scores to a common reference point across respondents. Approaches presented at the Sawtooth Software conferences have fused MaxDiff with rating scale data (Bacon *et al.* 2007, Magidson *et al.* 2009, Orme 2009). At the 2009 Sawtooth Software Conference, the inventor of MaxDiff, Jordan Louviere, mentioned a dual-response questioning device that he uses to deal with this issue when clients ask for MaxDiff scores with "absolute scaling." Louviere recently elucidated his approach via subsequent email correspondence with the author, and this article describes the idea and shows results of an empirical test. Louviere's dual-response approach is fortunately much more straightforward to implement than other solutions presented at previous Sawtooth Software conferences. The dual-response question can easily be added to any MaxDiff survey, and the scores may be estimated with standard MNL software (such as HB, logit, or latent class), using specialized coding of the design matrix (see Appendix for details).

Dual-Response Question

After each MaxDiff question, an additional question is posed to the respondent:

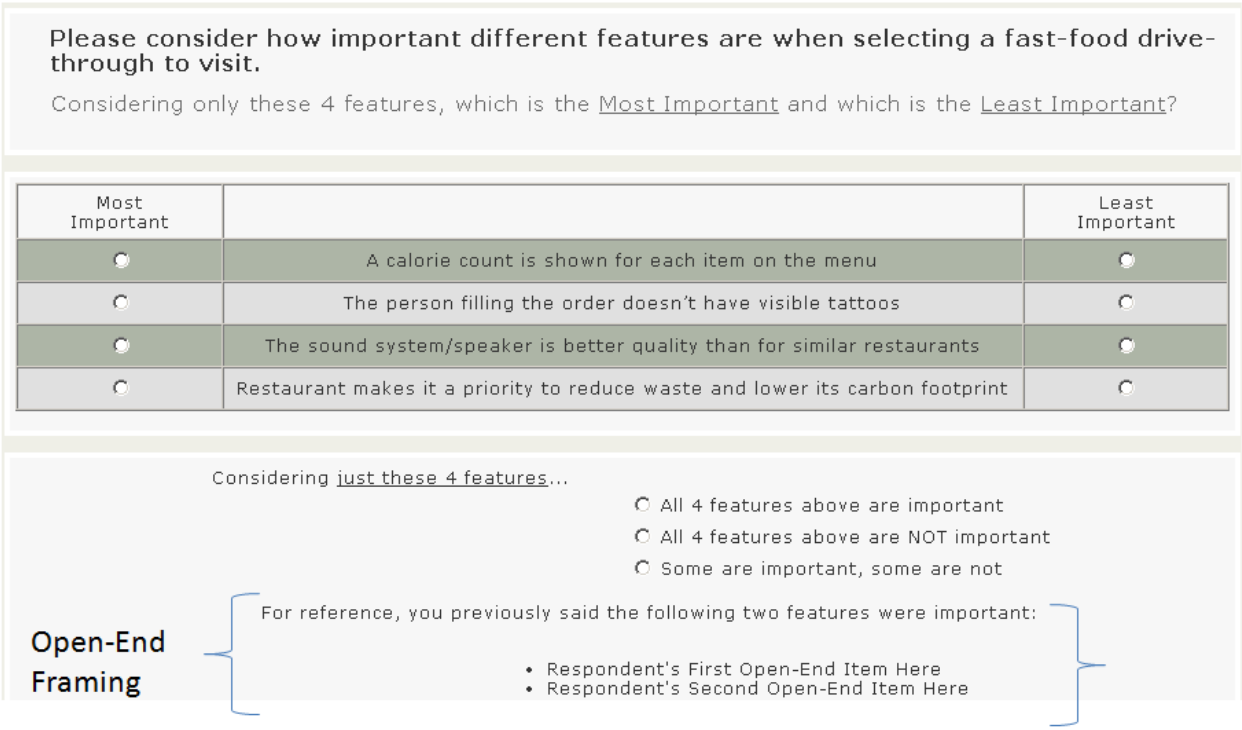
Considering just these four features…

- ❍ **All four are important**
- ❍ **None of these four are important**[1]
- ❍ **Some are important, some are not**

This question gives us the data to establish an anchor point for the scaling of items: the utility threshold between items deemed important vs. not important. Figure 1 shows how we implemented this in our empirical study involving features of fast-food restaurant drive-throughs.

[1] This wording was suggested by Rich Johnson, and is different from the wording we used in the experiment reported here.

Figure 1
MaxDiff Question with Dual-Response



Notice that Figure 1 includes an optional text-only section at the very bottom of the screen that we have labeled *Open-End Framing*. In previous research (Orme 2009), the author has found that asking respondents in an open-end fashion what items are important, and showing these responses in subsequent tasks involving the items the researcher brings to the study, provides important frame of reference that can lead to more reliable and concrete use of the item ratings.

Our inclination is to use four items per set when adding the dual-response question rather than five or more. The likelihood that the third option (some are important, some are not) is chosen increases as more items are shown per set. We would like to have a more even distribution across the three options to establish a more robust anchor point for our scaling.

Empirical Study Description

In June, 2009, we fielded a web-based study using Western Wats Opinion Outpost panel. The subject matter was features of drive-throughs at fast-food restaurants, and respondents were screened for frequent use of drive throughs.

The study involved two waves of data collection. In the first wave, we included 24 features having to do with the drive-through experience (MaxDiff in 12 sets, with each set featuring 4 features). 111 respondents completed the MaxDiff questionnaire, and we estimated item scores using aggregate logit. Based on the average scores for the sample, we divided the 24 features into two groups: the 12 that had the highest scores on average (stars), and the 12 that had the lowest scores on average (dogs).

We fielded a wave two questionnaire a week after fielding wave one, drawing the sample in the same way as before. This time, respondents received a MaxDiff survey (9 sets, again with 4 items per set) containing *either* the 12 stars or the 12 dogs. This design places standard MaxDiff at its greatest peril. Since it only obtains relative scores, there would be no way to know that the 12 stars were *all* more important (on average) than the 12 dogs. By using the dual-response approach, we hoped to anchor the scales to a common point and thus recover the original rank-order of importance found in wave one by using only wave two data. We also varied whether respondents were asked an open-ended question that was used to establish frame of reference. So, the four design cells in wave 2, and the sample sizes for each were:

Cell 1: 12 dogs, open-ended frame of reference (n=54)
Cell 2: 12 stars, open-ended frame of reference (n=54)
Cell 3: 12 dogs, NO open-ended frame of reference (n=57)
Cell 4: 12 stars, NO open-ended frame of reference (n=60)

This design approach is nearly identical to that used in previous MaxDiff experiments (also to try to resolve the relative scaling issue) by the author (Orme 2009).

Empirical Study Results

The median time to complete the first nine out of the twelve MaxDiff tasks in Wave 1 was 204 seconds (11.3 seconds per click, n=224[2]). The median time to complete nine dual-response MaxDiff tasks in Wave 2 was 232 seconds (n=225). This is a between-respondents comparison, so it doesn't control for between-respondent variation in response time. Given that caveat, asking the dual-response question appears to have added just 14% to the length of the task, or about 3 seconds to provide the dual-response answer (232-204)/9 tasks = 3.1 seconds.

The distribution of dual-responses across all 9 tasks for the different design cells is given in Table 1.

Table 1
Distribution of Dual-Responses

	Cells 1&3 (Dogs)	Cells 2&4 (Stars)
All 4 features are important	16%	44%
All 4 features are NOT important	12%	3%
Some are important, some are not	73%	53%

As expected, respondents who saw the 12 best items (stars) were much more likely to say that all four items presented in the MaxDiff task were important (44% vs. 16%). Strong differences in these judgments are critical to separating the stars from the dogs when developing the scores using just wave 2 data.

[2] Due to an error in programming a section of the wave 1 questionnaire (not related to MaxDiff), we repeated fieldwork for the wave 1 questionnaire the following week. Thus, we have 224 total respondents who completed 9 MaxDiff tasks without the dual-response question.

Half of the respondents (those in Cells 1 and 2) received an open-ended question prior to seeing the MaxDiff task (see Figure 2).

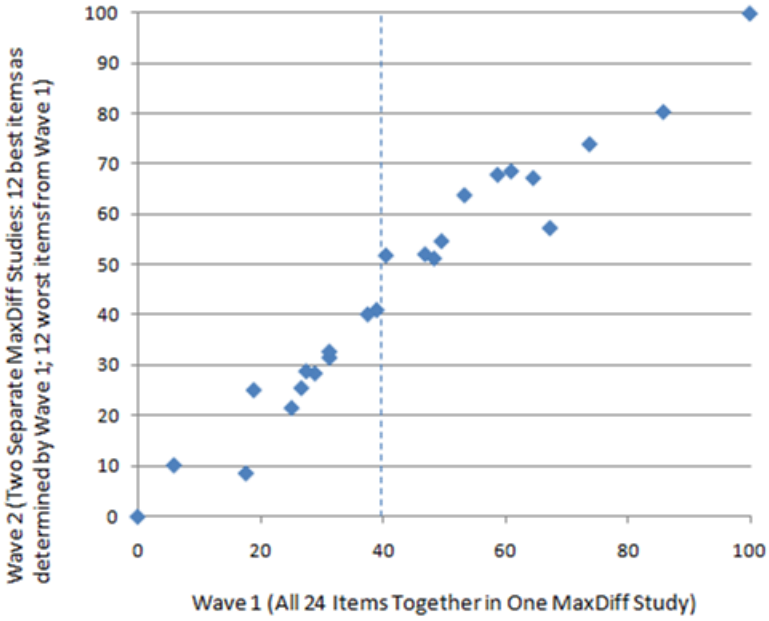
Figure 2
Open-End Question

The open-end questions took a median time to complete of 28 seconds. But, this is not wasted effort. Having the open-ended data can help the researcher examine (*post mortem*) whether important items were excluded from the MaxDiff study. Plus, asking the open-end (and reminding respondents of their answers) can be useful for establishing a frame of reference for respondents for judging the items in the questionnaire (Orme 2009), and possibly also increasing respondent engagement in the task.

Scores from wave two were estimated not on a relative scale (with arbitrary intercept, such as zero-centered), but using an Important/Not Important anchor point (contrained to be zero, via dummy-coding). Items that are considered important to the respondent are positive, and those considered not important are negative. The coding and estimation of the scores is described in the Appendix.

Figure 3 plots the original scores (on the x-axis) from wave 1 versus the scores estimated from the two separate MaxDiff studies collected in wave 2 (cells 1 and 2, which received open-end framing). The dotted line represents the delineation between the stars (top 12 items from wave 1) and the dogs (bottom 12 items from wave 1). To make it easier to present the data, we rescaled the scores from both waves to range from 0 to 100. For wave 2, we combined the raw MNL scores (since each set of respondents provided data only on half of the items) and then rescaled the full set of scores to run from 0 to 100.

Figure 3
Recovery of Wave 1 Scores Using Only Wave 2 Data
(With Open-End Framing)

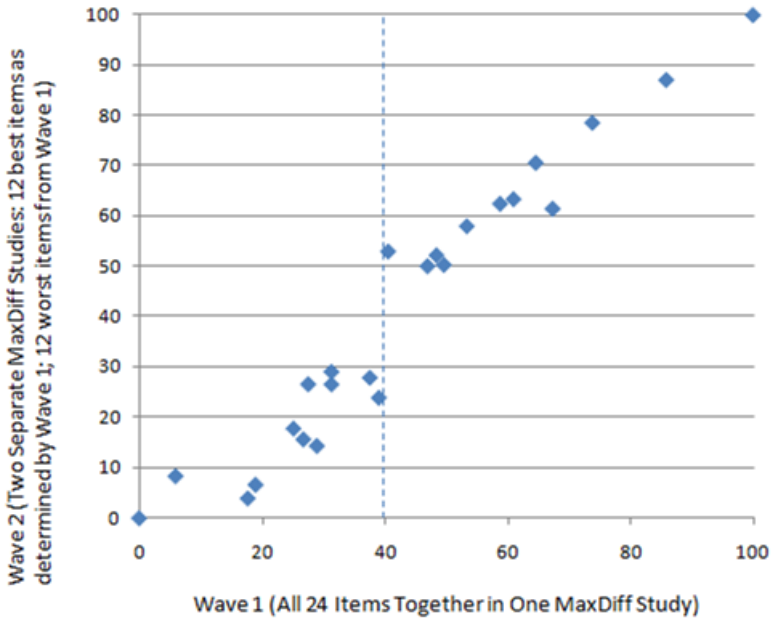


All twelve stars (as estimated from wave 1) are scaled above all twelve dogs using only the data from the two separate groups of respondents in wave 2. The R-Squared for the regression line fitting these points is 0.95. This exceeds the best fit (0.91) obtained in previous tests the author conducted using other approaches[3] to this same problem (Orme 2009). And, what is even more impressive, the dual-response approach has quite faithfully recovered the original scores from wave 1 using relatively small sample sizes (n=54 for both cells of wave 2).

Figure 4 shows the same plot, but using respondents from cells 3 and 4 on the y-axis (respondents who didn't get the open-ended question or the frame of reference reminder at the bottom of the dual-response question, see Figure 1).

[3] The two methods used in the previous research were calibration based on a 5-point rating scale, and a technique that used open-end supplied items and the MaxDiff items within additional paired-comparisons that were added to the MaxDiff design matrix.

Figure 4
Recovery of Wave 1 Scores Using Only Wave 2 Data
(Without Open-End Framing)



The R-Squared for the data in Figure 4 is 0.94. These respondents (cells 3 and 4) are nearly as successful at recovering the original scaling of all 24 items from wave 1 as respondents in cells 1 and 2. But, there is a discontiguous gap in the scores between the stars and the dogs for these respondents. Also, the scores for the 12 dogs are not nearly so correlated with wave 1 data as the scores for the 12 stars. A further examination of the data (described below) suggests that this may be due to relatively larger response error for respondents in cell 3. Larger response error leads to smaller differences in scores (reduced scale factor), in addition to the lower signal-to-noise ratio.

It would seem likely that respondents receiving the 12 dogs in wave 2 might answer with more noise (response error) than those receiving the 12 stars. The items on average are less engaging and important to them, so they naturally would be less enthusiastic about demonstrating their preferences among such generally weak features. We performed individual-level analysis with HB to further investigate the issue of response error for these wave 2 respondents.

HB provides a measure of fit (root likelihood) for each respondent that describes how well each respondent's scores fit the respondent's answers to the MaxDiff questionnaire. If the scores perfectly predict the choices, then the fit is 1.0. Random choices among the four items in the MaxDiff exercise would lead to a null fit of 0.25.

We hypothesized that two factors that influence the fit in the MaxDiff questionnaire for respondents in wave 2 were whether they received the 12 stars or dogs, and whether they got open-ended framing. To investigate this, we estimated a multiple regression, with two dummy-coded predictors: Got_Stars and Got_OpenEndedFraming.

First, we used the fit statistic from standard HB estimation of scores without the dual-response question as the dependent variable. The regression weights show that the fit statistic is 0.019 points higher when open-end framing is provided (t=1.3), and the fit is 0.044 points higher (t= 3.1) for respondents receiving the 12 best items instead of the 12 worst items.

We also computed scores via HB augmented by the information provided by the dual-response question (as described in the Appendix). The findings were similar, with fit 0.014 points higher (t=0.88) when open-end framing is provided, and 0.079 points higher (t=5.0) when respondents are evaluating the 12 best items instead of the 12 worst items.

So, the theory that the scores for the 12 dogs are compressed somewhat and noisy due to response error is supported. The fit is lower (error is higher) when respondents are evaluating the 12 dogs. And, the frame of reference technique may offer a small improvement in the level of engagement and fit to the data (this finding less conclusive, so we hope additional research can verify this point).

Summary and Conclusion

Jordan Louviere's dual-response task seems to offer an efficient and effective way to scale the scores relative to a common anchor point (in this case, the threshold utility between important and unimportant). Obtaining scores that are scaled relative to a common anchor point provides more information to characterize respondents than typical MaxDiff, where the scores have an arbitrary origin (and are often zero-centered). For example, with a common important/not important anchor point, one can see whether the items are all important or all not important for a respondent. This may lead to stronger comparisons between respondents and groups of respondents, and more meaningful segmentation results. Further research needs to be done to verify that the additional information provided by anchored scaling is not negated by error introduced because respondents are unreliable regarding their use of the dual-response to indicate that items are above or below the threshold of importance.

We should stop short of claiming that we now can obtain *absolute* scores from MaxDiff using the dual-response approach. The title of this paper purposefully uses the term *anchored*, which is less ambitious a descriptor than *absolute*. To state that the anchored scores are absolute is surely overstated, since response error directly affects the magnitude of the scores, and is a large source of between-respondent heterogeneity as estimated via HB (Islam *et al.* 2009). Furthermore, it can easily be argued that one respondent's threshold of importance may not necessarily map to another respondent's threshold, and there is a healthy degree of individual relativism involved in establishing an importance threshold, not to mention cultural/language differences between people that may affect the propensity to declare items important or not important. Thus, the promise of obtaining anchored scaling in MaxDiff comes at the cost of losing MaxDiff's robust property of avoiding bias in the scores due to difference in interpreting the meaning of scale anchors.

References

Bacon, Lynd, Peter Lenk, Katya Seryakova, and Ellen Veccia (2007), "Making MaxDiff More Informative: Statistical Data Fusion by way of Latent Variable Modeling," Sawtooth Software Conference Proceedings, pp 327-344.

Bacon, Lynd, Peter Lenk, Katya Seryakova, and Ellen Veccia (2008), "Comparing Apples to Oranges," Marketing Research, American Marketing Association, pp 29-34.

Islam, Towhidul, Jordan Louviere, and David Pihlens, (2009), "Aggregate choice and individual models: A comparison of top-down and bottom-up approaches," Sawtooth Software Conference Proceedings, Forthcoming.

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

Magidson, Jay, Dave Thomas, and Jeroen K. Vermunt (2009), "A New Model for the Fusion of MaxDiff Scaling and Ratings Data," Sawtooth Software Conference Proceedings, Forthcoming.

Orme, Bryan (2009), "Using Calibration Questions to Obtain Absolute Scaling in MaxDiff," SKIM/Sawtooth Software European Conference, Prague, Czech Republic.

Appendix[4]

Coding Dual-Response MaxDiff Designs

Consider a twelve-item MaxDiff study, where the respondent sees items 1, 4, 7, and 12 in the first task, and indicates that 1 is best and 12 is worst. The standard coding would use k-1 columns in the design matrix (to avoid linear dependency). Sawtooth Software's standard practice is to use dummy-coding for MaxDiff experiments, and to omit the last level as the reference level.

The typical MaxDiff coding (following Louviere's early MaxDiff papers) leads to the following design matrix for the "best" task and the "worst" task:

"Best" task design matrix:

1	0	0	0	0	0	0	0	0	0	0	(this row selected)
0	0	0	1	0	0	0	0	0	0	0	
0	0	0	0	0	0	1	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	

"Worst" task design matrix:

-1	0	0	0	0	0	0	0	0	0	0	
0	0	0	-1	0	0	0	0	0	0	0	
0	0	0	0	0	0	-1	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	(this row selected)

Consider a dual-response MaxDiff questionnaire. To apply anchored scaling (with the anchor utility at zero), we introduce a new item to the design matrix (importance threshold) and we treat it as if it were the 13th item (the reference, zero item). Now, we code the design matrix with 12 columns, where the 12th item from before is no longer the reference item (so it gets coded as a 1 in the 12th column):

"Best" task design matrix (modified to have 12 columns):

1	0	0	0	0	0	0	0	0	0	0	0	(this row selected)
0	0	0	1	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	1	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	1	

[4] The original 2009 version of this paper included an incomplete coding procedure in this Appendix that had deficiencies, especially for HB estimation. The coding procedure didn't insert the anchor item within the "best" and "worst" tasks if respondents indicated "some are important, some are not." No action was taken in this case. We thank Paul Johnson of Western Wats for suggesting that the anchor item should be inserted in the "best" and "worst" tasks to properly account for this condition.

"Worst" task design matrix (modified to have 12 columns):

```
-1   0   0   0   0   0   0   0   0   0   0   0
 0   0   0  -1   0   0   0   0   0   0   0   0
 0   0   0   0   0   0  -1   0   0   0   0   0
 0   0   0   0   0   0   0   0   0   0   0  -1      (this row selected)
```

Next, we describe how to modify the data file to code Louviere's follow-up question.

If the respondent indicates for this task that "some items are important, some are not," then we modify the "best" and "worst" tasks as if the respondent also saw the 13th (anchor threshold) item in both tasks, but didn't think it was either the best or worst item.

"Best" task design matrix (modified to insert the threshold item):

```
 1   0   0   0   0   0   0   0   0   0   0   0      (this row selected)
 0   0   0   1   0   0   0   0   0   0   0   0
 0   0   0   0   0   0   1   0   0   0   0   0
 0   0   0   0   0   0   0   0   0   0   0   1
 0   0   0   0   0   0   0   0   0   0   0   0      (inserted threshold item)
```

"Worst" task design matrix (modified to insert the threshold item):

```
-1   0   0   0   0   0   0   0   0   0   0   0
 0   0   0  -1   0   0   0   0   0   0   0   0
 0   0   0   0   0   0  -1   0   0   0   0   0
 0   0   0   0   0   0   0   0   0   0   0  -1      (this row selected)
 0   0   0   0   0   0   0   0   0   0   0   0      (inserted threshold item)
```

If the respondent indicates (after choosing item 1 as best and item 12 as worst) that <u>all four items are important</u>, we retain the original "best" and "worst" tasks with 12 columns and 4 alternatives. We don't insert the threshold item into the "best" and "worst" tasks as shown directly above. Instead, we augment the data file with an additional task to account for the dual-response (with five alternatives to include the 13th threshold item as an alternative) that indicates that all four items are preferred to the 13th item (this can be done by formulating another "worst" task, where the dummy codes are inverted and the 13th item is selected as "worst".):

Augmented task to account for all items deemed important:

```
-1   0   0   0   0   0   0   0   0   0   0   0
 0   0   0  -1   0   0   0   0   0   0   0   0
 0   0   0   0   0   0  -1   0   0   0   0   0
 0   0   0   0   0   0   0   0   0   0   0  -1
 0   0   0   0   0   0   0   0   0   0   0   0      (this row selected)
```

If the respondent indicates that <u>all four items are NOT important</u>, again we retain the original "best" and "worst" tasks with 12 columns and 4 alternatives. We add a new task where the 13th threshold item is viewed to be preferred to the other items within the set. That augmented task looks like:

Augmented task to account for all items deemed NOT important:

1	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	1	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	1	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	1	
0	0	0	0	0	0	0	0	0	0	0	0	(this row selected)

Note: it is also possible to code the augmented tasks as a series of paired comparisons. For example, if the respondent indicates that all items in a set are important, then each item in the set is paired vs. the 13th threshold item, and chosen. We used this procedure for the calculations in this paper, though we have tried coding it both ways and the method above (which is more compact) gives essentially identical results.