Using Calibration Questions to Obtain Absolute Scaling in MaxDiff

(presented at the 2009 SKIM/Sawtooth Software European Conference)

Bryan Orme[1], Sawtooth Software

March, 2009

Background

MaxDiff (Maximum Difference Scaling) has become a widely-used survey research tool for scaling items in terms of preference or importance. The methodology was developed by Jordan Louviere in the late 1980s (Louviere 1991, Finn and Louviere 1992). For an introduction to MaxDiff, we recommend the "MaxDiff/Web Technical Paper" (Sawtooth Software 2007). MaxDiff leads to robust scaling of items, with much higher precision than standard rating scales (Cohen and Orme 2004). It avoids scale use bias, making it particularly useful for cross-cultural research. Despite the many benefits of MaxDiff, one weakness is that items are placed only on a relative (rather than an absolute) scale. Since respondents tell us only which item is best and worst in each set, we do not learn anything about the *absolute* desirability or importance of the items. A respondent who tends to dislike all the items in a MaxDiff study cannot be distinguished from a respondent who tends to like all the items.

How problematic is this lack of absolute scaling for MaxDiff? If a wide variety of items are included in the experiment, representing essentially the full spectrum of possible bad items to good items, then the problem certainly is reduced. Also, if an item that has common preference or importance across respondents is made the reference point, we can scale the data with respect to this reference (but finding such an item is much easier said than done). Added to these issues, the fact that the scale factor (response error) may differ significantly between respondents can further complicate our ability to compare respondents on the derived scores.

The lack of absolute scaling with MaxDiff raises the question as to whether we can consistently make accurate comparisons between respondents or groups of respondents on MaxDiff scores[2]. Although MaxDiff is considered an excellent methodology for segmenting respondents into groups with divergent preferences, it is possible that it could do even better absent the relative scaling issue.

Abstract

In this paper, we create an artificial situation that demonstrates the relative scaling issue for MaxDiff at its worst. We collect a first wave of MaxDiff data on 30 items, and based on the items' average scores we separate them into the best 15 and the worst 15 items. Then, we give two new sets of respondents MaxDiff questionnaires that include *either* the best 15 or the worst 15 items as determined from Wave 1 (plus a few calibration questions). With the addition of the calibration questions, we attempt to recover the original scaling of the first 30 items *using only the data from wave 2*. The two types of calibration questions involve a subset of five MaxDiff items judged using a 5-point rating scale, or the method of paired comparisons (MPC) involving best and worst items volunteered by the respondents (via open-end questions) versus a subset of five items included in MaxDiff. The 5-point scale calibration data allow us to rescale Wave 2 data and fit the original scaling of Wave 1 scores with R-Squared of up to 0.83. We find even greater success in a quali-quantitative calibration approach. A set of paired comparisons

[1] The author thanks Rich Johnson and Lynd Bacon for their critique of earlier drafts. The opinions and any errors are the sole responsibility of the author.

[2] Many other measures used in market research also are subject to relative scaling problems, including constant-sum, ranking, and potentially even the standard 5-point Likert scale.

involving volunteered absolute best and worst items compared to a subset of items from the MaxDiff questionnaire seems to work well for establishing respondent-specific framing and aggregate recovery of the original relationship among the 30 items as measured in the first wave (R-Squared 0.91). A second study is reported that confirms findings of the first study.

Wave 1 Data Collection

We fielded two waves of data collection (using Western Wats' Opinion Outpost online panel) on successive days in the week just prior to the 2008 presidential elections (October 28-29, 2008). This was a period of great anxiety for the US public, as the stock market had cratered, banks had failed, and the housing and credit crisis seemed at a peak. Barack Obama was leading in the polls over John McCain with one week to go.

For Wave 1, we studied 30 issues/policies that politicians might focus on or try to enact. In previous research into political attitudes, we had seen vastly different opinions between Republicans and Democrats (Orme and King 2008). To reduce heterogeneity of the preferences for our experiment, we decided to focus only on respondents that generally affiliated with the Democratic party. After deleting the 10% fastest responders, we were left with 150 completed interviews in Wave 1. Each respondent received 18 MaxDiff questions, each showing 5 items per set (see Appendix A for example question).

We computed the scores using aggregate logit, and rescaled the scores so the worst item was 0 and the best 100 (Table 1). The higher the score, the more desirable the issue/policy, on average, for this group of Democratic-leaning respondents.

Table 1
MaxDiff Scores for 30 Items
(Normalized: Worst Score = 0 and Best Score = 100)

Score	Item
100	Reduce taxes for middle and lower income households
93	Guarantee national health care and elder care program
93	Ensure the long-term health of Social Security
93	Enact policies to improve general economic climate and create jobs
86	Develop alternative energy sources
81	Reduce our reliance on foreign oil imports
75	Increase funding for education
74	Enact policies to solve housing / mortgage crisis
74	Reduce the federal deficit
72	Reduce US troop involvement in Iraq
60	Reduce corruption / Improve ethics in government
59	Create a national jobs program
59	Increase funding to help homeless / hungry
56	Restrict carbon emissions to reduce global warming
52	Improve food safety and increase food supply
46	Improve our relations / reputation with other countries
45	Improve infrastructure such as roads and rails
43	Reduce trade deficit with foreign countries
40	Reduce illegal immigration
37	Strengthen women's reproductive "Right to choose"
33	Improve race relations
24	Increase worldwide humanitarian efforts
22	Enact campaign finance reform
17	Reduce illicit drug use
17	Impose term limits for Congress
10	Increase US troop involvement in Afghanistan
6	Give full marriage rights to gays
6	Increase defense / military spending
2	Restrict gun ownership
0	Increase spending in the war on terrorism

Based on these data, we divided the items into two lists: the best 15 items and the worst 15 items. The next day, we interviewed a second wave of respondents. These respondents received MaxDiff questionnaires including only 15 items: either the 15 best or 15 worst items.

Wave 2 Data Collection

Respondents in Wave 2 received either the 15 best items or 15 worst items, on average, as determined by the first wave of respondents interviewed the day before (respondents were prohibited from participating in both waves). The sample was drawn in the same way by Western Wats as Wave 1. Rather than use a standard MaxDiff questionnaire in Wave 2, we employed an adaptive variety of MaxDiff. The author has shown that the Adaptive MaxDiff methodology produces mean scores nearly identical to standard MaxDiff (Orme 2006). With Adaptive MaxDiff, whenever items are marked "worst," they are discarded from further consideration. The questions continue in multiple rounds (as in rounds of a round-robin tournament) until an overall winning item is identified (see Appendix B for adaptive design). It took five rounds to identify the winning item in a 15-item experiment.

Wave 2 respondents were randomly divided into six groups, each receiving a slightly different treatment and version of the questionnaire.

Calibration via 5-Point Ratings Scale

Two groups of respondents in Wave 2 were randomly selected to receive an Adaptive MaxDiff questionnaire followed by a calibration ratings grid. After deleting the 10% fastest responders, the sample sizes for these two cells were 115 (receiving worst 15 items) and 96 (receiving best 15 items).

In the calibration grid, respondents were asked to rate five of the items on a 5-point desirability scale. We used a radio-button grid, with scale points labeled as follows:

1	2	3	4	5
Extremely Bad	Bad	Good	Very Good	Extremely Good
❍	❍	❍	❍	❍

The five items rated were customized based on each respondent's answers to the Adaptive MaxDiff questionnaire:

Item1: Item winning Adaptive MaxDiff tournament
Item2: Item not eliminated until 4th Adaptive MaxDiff round
Item3: Item not eliminated until 3rd Adaptive MaxDiff round
Item4: Item not eliminated until 2nd Adaptive MaxDiff round
Item5: Item eliminated in 1st Adaptive MaxDiff round

The average ratings on the 5-point scale for customized items included in the ratings grid were:

Table 2: Mean Ratings on 5-Pt Scale

	N=115 Worst15	N=96 Best15
Winning MaxDiff Item	3.99	3.98
Item Eliminated in 1st Round	2.30	3.16

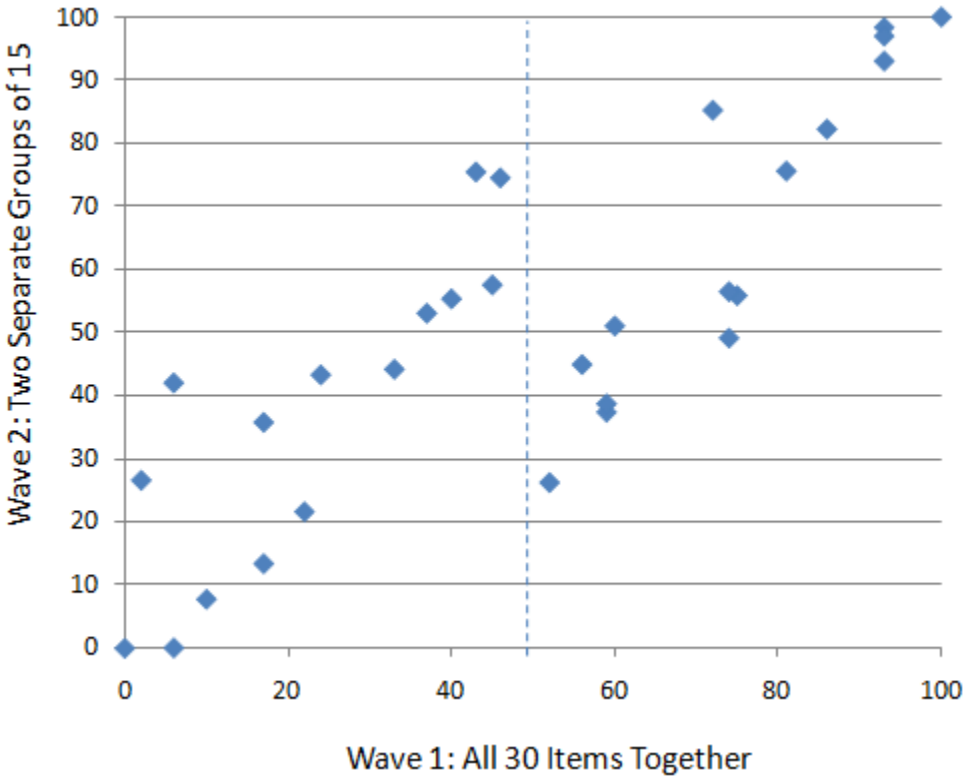
As expected, respondents' scores for the item they found best within their set of 15 are higher than items they eliminated in the 1st round of the Adaptive MaxDiff experiment. However, it is concerning that the average ratings that respondents give to the best item they saw in their respective sets of 15 items are essentially tied. We would expect that the rating for the winning item should be higher among the respondents who saw only the best 15 items on average as determined in Wave 1. It is likely that respondents tend to use the 5-point rating scale in a somewhat relative sense, adjusting their ratings within the context of items seen in the questionnaire. This would make it difficult to use the ratings data as an absolute measuring stick to calibrate the Wave 2 Adaptive MaxDiff scores and recover the pattern of scores seen from Wave 1. Even so, we made the attempt.

We used a rather simple method to develop a common reference point for scaling the two sets of Adaptive MaxDiff data. We augmented the Adaptive MaxDiff data with additional implied paired comparisons derived from the ratings on the 5-point scale. We examined the distribution of responses to the 5-point scale (pooling ratings on all five items), and chose a division between points on the scale to represent an

arbitrary, yet common, threshold of desirability. The division between scale points 3 and 4 on the 5-point scale was the best cut for dividing the distribution roughly in half. That division can represent a threshold of desirability. We add the threshold as an additional item within the data matrix (the reference item) and constrain it to zero. For each respondent, the five MaxDiff items were either "chosen" or "rejected" with respect to the threshold item, depending on the observed ratings. The threshold, of course, serves as the common reference for scaling between our two respondent groups receiving either the best 15 or worst 15 items.

We computed the scores separately for the two groups of respondents using aggregate logit. We combined the two sets of raw scores, and rescaled the combined 30 items so the worst item was 0 and the best 100. The rescaled scores are plotted versus the scores from Wave 1 in Figure 1.

Figure 1
Rescaled Scores: Wave 1 vs. Wave 2 Respondents
(Calibration Based on 5-pt Rating Scale)



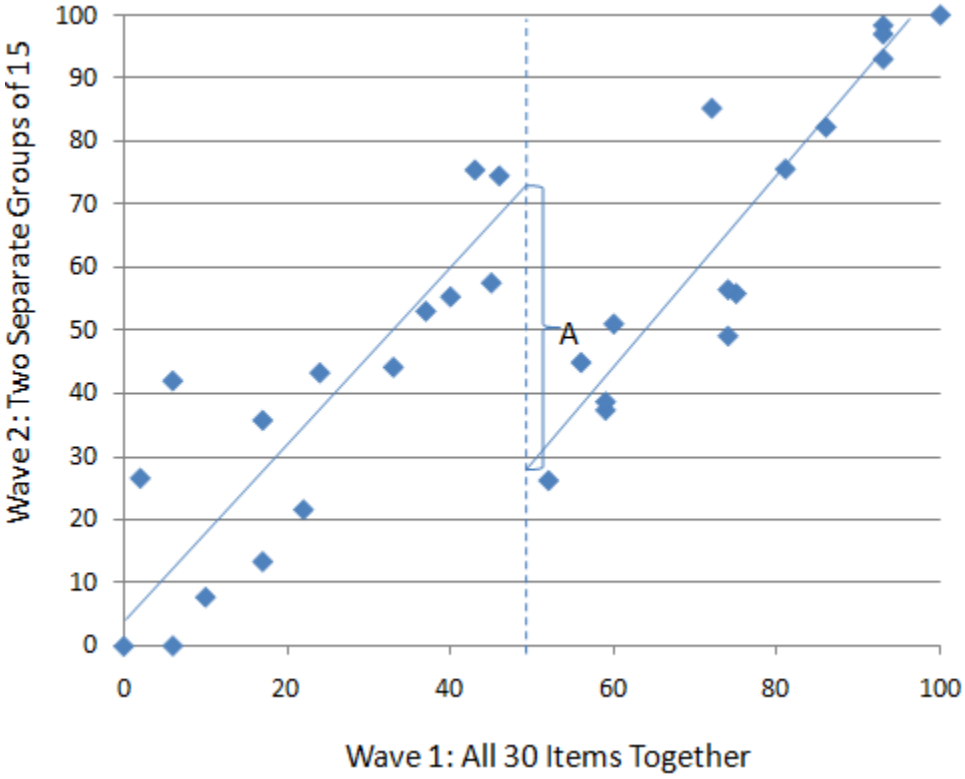
The dotted line in the chart marks the boundary between the worst and best 15 items from Wave 1. The R-Squared for the regression line fitting the data is 0.70. Of course, we shouldn't expect to achieve perfect fit due to sampling error[3], since different respondents were used in waves 1 and 2. (As a point of reference, we randomly split the original 150 respondents from wave 1 into two equal replicates, estimated scores separately for the replicates, and computed the R-Squared between the two sets of

[3] There are two additional sources of error: 1) There is a slight variance in methodology: standard MaxDiff and Adaptive MaxDiff, which could lead to subtle differences; 2) The two respondent groups may have exhibited different amounts of response error leading to different scale factors in the raw scores, which were simply combined.

scores. Across repeated random split samples, the average R-Squared was 0.96. This represents a high-water mark for possible fit[4].)

Although the calibrated scores from Wave 2 seem to loosely fit the data from Wave 1, Wave 2's data seem to have a vertical shift between the best 15 and worst 15 items that reflects fundamental difficulty in matching the two sets of scores. Figure 2 illustrates that shift (quantity A) after fitting lines to the two separate halves of the data.

Figure 2
Rescaled Scores: Wave 1 vs. Wave 2 Respondents
(Calibration Based on 5-pt Rating Scale)



Calibration via 5-Point Ratings Scale with Elicited Frame of Reference

Two additional groups of respondents in Wave 2 were given a slightly different calibration exercise. Prior to receiving any MaxDiff questions, we asked them to volunteer (via open-end response) the two absolute best items they felt were most important that our leaders accomplish or concentrate on. Then, we asked them to volunteer the two absolute worst items they felt our leaders should accomplish or concentrate on (see Appendix A for the text of the questions).

After all 15 items had been introduced to respondents in the Adaptive MaxDiff questionnaire (after round 1), we displayed the volunteered 4 items to the screen and allowed the respondents to review and edit their responses if desired. We expressed to them that it was important that their answers be complete and

[4] We recognize that the hypothetical "high-water mark" is slightly understated for this example, since the sample sizes (n=115 and n=96) are a bit bigger than the 75 as were used in the repeated sampling simulations with Wave 1 data.

that they be happy with them, or a later stage of the questionnaire wouldn't work properly (see Appendix A for the text of these questions).

After the Adaptive MaxDiff survey, respondents received the 5-point rating scale for calibration purposes as described before. But, these respondents were also asked to rate their open-end items within the grid, directly above the five customized items from the MaxDiff survey. We were hoping that adding these four customized best and worst items to the grid would lead to a better frame of reference so that respondents could better use the 5-point scale in an absolute sense to convey degree of preference for the items included in the questionnaire.

After deleting the 10% fastest responders, and deleting respondents who could not provide four legitimate answers to the open-end questions (15% were deleted due to incomplete or inappropriate open-end answers), the sample sizes for these two cells were 96 (receiving worst 15 items) and 86 (best 15 items).

The average ratings on the 5-point scale for customized items included in the grid were:

Table 3: Mean Ratings on 5-Pt Scale

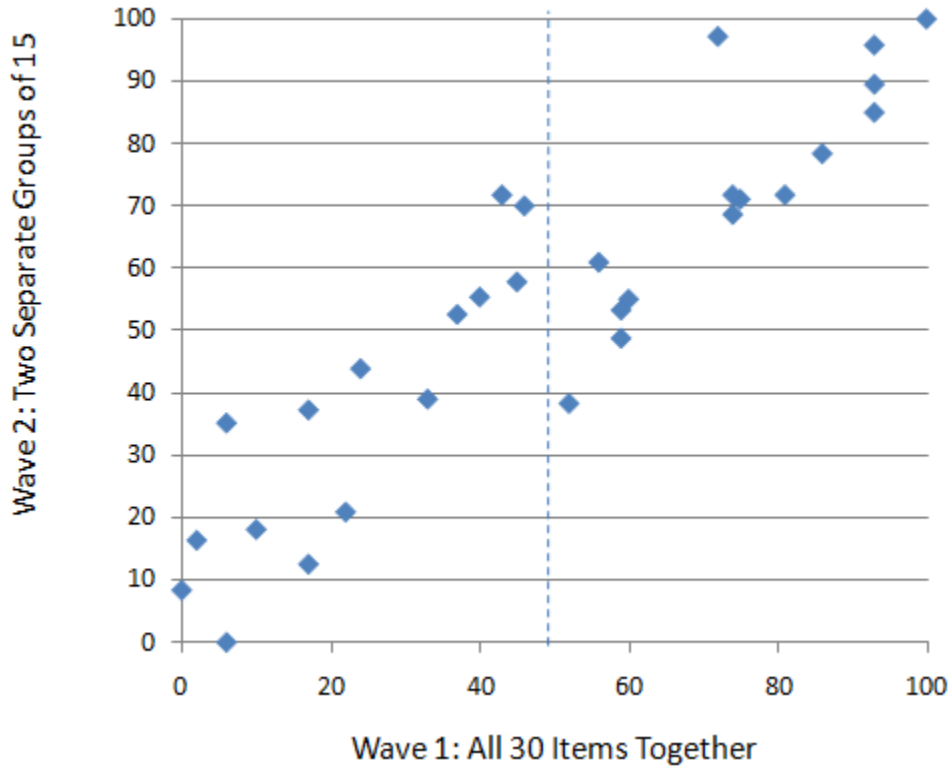
	N=96 Worst15	N=86 Best15
Winning MaxDiff Item	3.64	3.94
Item Eliminated in 1st Round	2.13	2.95

These data look a bit more differentiated than those from the respondents who didn't rate volunteered (open-end) items within the grid (compare to Table 2). It seems that asking respondents to rate their volunteered absolute best and worst items prior to the five items we included in the questionnaire provided a better frame of reference so that 5-point rating scale could be used in a relatively more absolute sense.

We again examined the distribution of 5-point ratings across the five items in the calibration section, again finding the appropriate cut point between points 3 and 4 on the scale. We augmented the Adaptive MaxDiff data with five paired comparison questions as described earlier. The ratings for the volunteered items were not included in the score estimation. The raw logit scores were again combined, then rescaled within the range 0-100.

Figure 3 shows the results for these Wave 2 respondents plotted against the scores from Wave 1.

Figure 3
Rescaled Scores: Wave 1 vs. Wave 2 Respondents
(Calibration Based on 5-pt Rating Scale, with Volunteered Item Framing)



The R-Squared for the regression line fitting the data is 0.83. Although there appears to remain some shift in scores from the worst 15 items relative to the best 15 items, the problem is less pronounced than when respondents didn't first rate elicited absolute best and worst items as reference. The R-Squared is improved over the 0.70 when the calibration didn't include volunteered reference items.

We conclude from these two calibration exercises that we can do a reasonable job recovering the original scaling of Wave 1 scores with calibration via the 5-point scale on a customized subset of the items. But, the calibration works better if respondents are also asked to rate their elicited absolute best and worst items as reference. In the next section, we report on another calibration method that worked even better and may be more solid from a methodological and statistical standpoint.

Paired Comparison Calibration Results

Two additional groups of respondents in Wave 2 did not use a 5-point scale for calibration. Rather, they received eight additional paired-comparison judgments after completing the Adaptive MaxDiff questionnaire (see Appendix A for sample questions). These questions compared the respondents' volunteered open-end items directly to the 15 items used in the Adaptive MaxDiff questionnaire. Respondents were asked to choose which item in each pair they preferred our leaders accomplish or concentrate on (or they could specify that the two items essentially meant the same thing to them). The third option (tie) was necessary, because it is quite possible that respondents could volunteer an open-end item that was nearly identical to one included in the questionnaire. For this same reason, these calibration

tasks cannot be integrated within the MaxDiff tasks involving three or more items (how could respondents choose the best or worst within the set in the case of ties?)

The design of the pairs, as well as the percent of respondents who chose their open-end item as "best" in each pair are given in Table 4.

Table 4: Paired Comparison Calibration Tasks

	% Respondents Selecting Open-End (OE) Items as More Preferred than MaxDiff Items	
	Worst 15 N=80	Best 15 N=76
OE Best vs. Adaptive MaxDiff Winner	70%	25%
OE Best vs. 4th round survivor	93%	78%
OE 2nd Best vs. 4th round survivor	89%	68%
OE 2nd Best vs. 2nd round survivor	94%	88%
OE Near Worst vs. 4th round survivor	41%	7%
OE Near Worst vs. 2nd round survivor	43%	20%
OE Worst vs. 2nd round survivor	48%	18%
OE Worst vs. 1st round loser	58%	22%

We see quite large differences in the choices between respondent groups. As one example, respondents receiving the best 15 items felt their open-end best item was a better policy for leaders to work on 25% of the time. Respondents receiving the worst 15 items felt their open-end best item was a better policy 70% of the time. It would appear that these paired comparison data may give us a good opportunity to distinguish the absolute desirability of the items between the two groups.

The eight paired-comparison questions were added to the Adaptive MaxDiff data matrix, with the four open-end responses coded as four additional (common) items. Three additional synthetic pairs were added to the data: Open-End Best preferred to Open-End 2nd Best; Open-End 2nd Best preferred to Open-End Near Worst; Open-End Near Worst preferred to Open-End Worst. As with any MaxDiff estimation, one item is chosen as the reference item and constrained to zero.

We again used aggregate logit to compute the scores separately for cells 5 and 6. After estimating the raw scores, we rescaled the data *individually* for the two independent models so that the worst score was 0 and the best 100 within each group[5]. The results are shown in Tables 5 and 6 (and the original scores from Wave 1 are shown for reference).

[5] The fact that we can rescale the data for each model bounded by 0 and 100 helps avoid the scale use (response error) heterogeneity issue. The scale is normalized to have a 100-point range for each group of respondents, and we assume that absolute best and worst end-points have been included in the experiment for each individual.

Table 5: Scores for 15 Worst Items

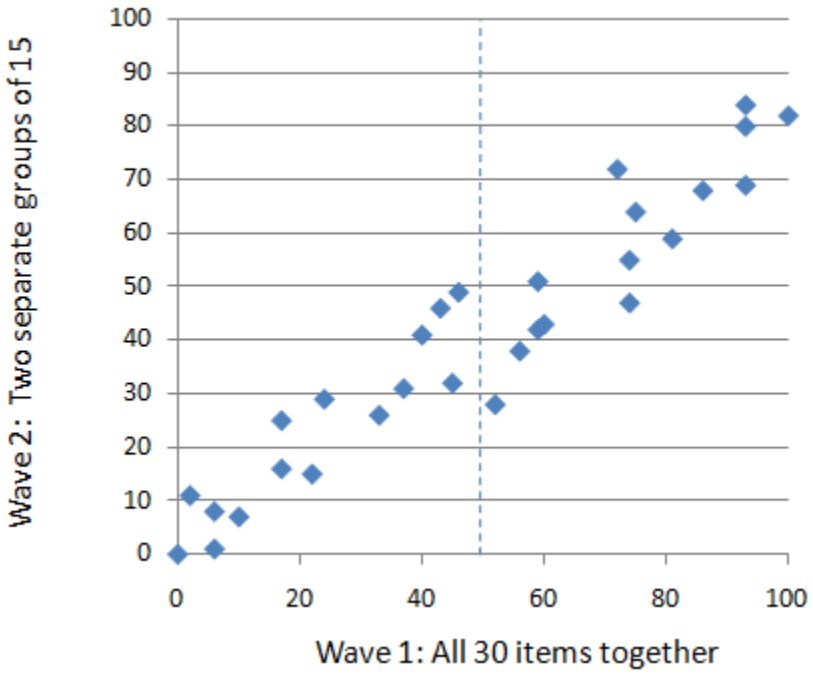
Wave 1 Score	Wave 2 Score	Worst 15 Items from Wave 1
N/A	100	Open-End Best
N/A	71	Open-End 2nd Best
46	49	Improve our relations / reputation with other countries
45	32	Improve infrastructure such as roads and rails
43	46	Reduce trade deficit with foreign countries
40	41	Reduce illegal immigration
37	31	Strengthen women's reproductive "Right to choose"
N/A	28	Open-End 2nd Worst
33	26	Improve race relations
24	29	Increase worldwide humanitarian efforts
22	15	Enact campaign finance reform
17	25	Reduce illicit drug use
17	16	Impose term limits for Congress
N/A	10	Open-End Absolute Worst
10	7	Increase US troop involvement in Afghanistan
6	8	Give full marriage rights to gays
6	1	Increase defense / military spending
2	11	Restrict gun ownership
0	0	Increase spending in the war on terrorism

Table 6: Scores for 15 Best Items

Wave 1 Score	Wave 2 Score	Best 15 Items from Wave 1
N/A	100	Open-End Best
N/A	86	Open-End 2nd Best
100	82	Reduce taxes for middle and lower income households
93	84	Guarantee national health care and elder care program
93	80	Enact policies to improve general economic climate and create jobs
93	69	Ensure the long-term health of Social Security
86	68	Develop alternative energy sources
81	59	Reduce our reliance on foreign oil imports
75	64	Increase funding for education
74	55	Reduce the federal deficit
74	47	Enact policies to solve housing / mortgage crisis
72	72	Reduce US troop involvement in Iraq
60	43	Reduce corruption / Improve ethics in government
59	51	Create a national jobs program
59	42	Increase funding to help homeless / hungry
56	38	Restrict carbon emissions to reduce global warming
N/A	29	Open-End 2nd Worst
52	28	Improve food safety and increase food supply
N/A	0	Open-End Absolute Worst

And, the scatter plot comparing Wave 1 scores to Wave 2 for all 30 items is given in Figure 4.

Figure 4: Wave 1 vs. Wave 2 Scores



The R-Squared for the regression line fitting the data is 0.91 (nearly reaching the theoretical ceiling of 0.96). For this project, it appears that the calibration via additional paired comparison judgments has done a creditable job of circumventing the relative scaling limitation of MaxDiff. This final method performs better than the other two calibration methods that employed 5-point ratings (with R-Squared of 0.83 and 0.70, respectively).

Note that the slope of the data in Figure 1 seems a bit flatter than 1.0. This is due to the fact that our 30 items didn't include the absolute best item for every respondent. Whereas a score of 100 is automatically given to the top item in the scaling of Wave 1 data on the X-axis, the items on the Y-axis are scaled with respect to a theoretically maximum item that was elicited from each respondent. Therefore, we shouldn't expect that the scores should lie on a 45-degree line that passes through 100. The degree to which this line becomes flatter depends on how well the set of items used in the experiment included both the extreme worst and best levels for the sample of respondents.

Strength of Calibration Approaches

Using paired comparison judgments in the calibration question avoids scale use bias. The fact that respondents use a 5-point scale differently was an additional (and significant) source of error in the models that involved augmenting the Adaptive MaxDiff data with contrived paired comparisons based on an arbitrary, yet fixed, cut point from a 5-point rating scale. The direct paired comparison judgments integrating the volunteered reference items in the calibration section are free from scale use bias and are congruent with other Adaptive MaxDiff questions employed in the survey (the 4th round in the Adaptive MaxDiff experiment involves paired comparisons, see Appendix B).

Although we have only shown aggregate analysis in this paper, these calibration methods can be employed within disaggregate analysis (latent class or HB). With disaggregate analysis, the benefits of avoiding idiosyncratic scale use bias (e.g. the 5-point scale) and employing calibration questions that

reflect similar context as the MaxDiff information would seem even more useful. Such calibration questions should also reflect more similar response error (and implied scale) as the Adaptive MaxDiff questions, relative to contrived paired comparisons derived from 5-point rating scales.

Open-End Items for Calibration Purposes

Using the quali-quantitative calibration methodology we've presented here involves adding open-ended questions to the survey. There are pros and cons to this addition. Aside from the additional time requirement, a concern is whether respondents will type appropriate and complete open-end information. After deleting the speeders (10% of the sample), we were left with 399 respondents who received open-end questions. The data below show what percent of respondents were able to provide four seemingly legitimate answers (not just "None" or "Don't Know" or stray characters filling the blank).

N=399

81%	Provided four valid issues/policies on first try
4%	Provided four valid issues/policies only upon prompting for second try
15%	Unable to provide four valid issues/policies

Although it is a subjective judgment, it seems good that 85% of respondents provided legitimate answers to four open-ended questions in an online survey. Furthermore, the open-end data have value in and of themselves. They can be examined to determine if important items were missing from the questionnaire. Although it leads to additional work to view the responses and manually clean the data, researchers are always looking for good ways to identify less motivated or less knowledgeable respondents. This would seem to provide an additional tool for that.

Another concern is whether asking open-ends annoys respondents. Prior to deleting the 15% of respondents who provided incomplete answers to the four open-end items, we tallied responses to the following question about respondents' attitude toward open-ends in online surveys.

> **Towards the beginning of this survey, we asked you to type some issues that were pertinent to you. Regarding open-end questions (answers that you are required to type), which best describes your opinion?**

(n=399)

43%	I LIKE it when surveys include questions where I am asked to type my answer
43%	It doesn't matter to me one way or the other
14%	I DISKLIKE it when surveys include questions where I am asked to type my answer

These results suggest that more respondents prefer giving open-end answers than are turned off by doing so, so the idea of eliciting open-end items for inclusion in calibration sections seems feasible.

Adaptive MaxDiff Necessary?

This research naturally raises the question of whether similar success could be found using the more common, standard MaxDiff rather than Adaptive MaxDiff. We believe the calibration will be more effective if the five MaxDiff items are chosen to span the largest degree of preference as possible for each respondent. Thus, it would seem to require a computer-administered design, where both highly preferred and highly disliked items are chosen in the calibration section to be compared to the open-end items. It

seems that this could be done using simple counting analysis and standard MaxDiff, but the Adaptive MaxDiff methodology naturally leads to identification of a suitable range of items to utilize in calibration.

Limitations

Rating scales suffer from scale use bias. There is no way to know that, say, a 3.0 for respondent A means the same thing as a 3.0 for respondent B. Our results show that respondents seem to adjust their ratings within the context of the items given them in the questionnaire. But, asking respondents to rate their volunteered open-end items alongside the items of interest appears to improve their ability to apply the scale in a (relatively) more absolute sense. Still, it seems that relying on rating scales to identify a common scale point to resolve the relative scaling issue in MaxDiff will remain difficult.

Although we achieved good results here with our final model, when respondents are asked to volunteer (open-ended) their extreme items, do they necessarily define the same absolute (and comparable) range of preference? To illustrate the potential problem, if I ask a respondent in Seattle, Washington to go outside and identify the tallest and shortest trees she can see, does this establish a common scale that I can use to compare to a respondent performing the same exercise but living in Waco, Texas[6]? The probable lack of ability for humans to invoke a common range via open-ended responses could add a significant source of error or bias to the model.

Study #2, Consumer Confidence

After seeing the positive results in our first study, we decided to conduct a second study to see if we could reproduce the favorable outcome. This time, we used the context of consumer confidence (a deep recession had gripped the US and abroad throughout 2008, and was formally declared by economists in Q4 2008). We included 30 items related to consumer confidence (such as unemployment rates, job security, interest rates, value of investments, real estate values, personal tax rates, affordability of food, etc.). The MaxDiff questions were posed in terms of importance rather than desirability (see Appendix C for example question).

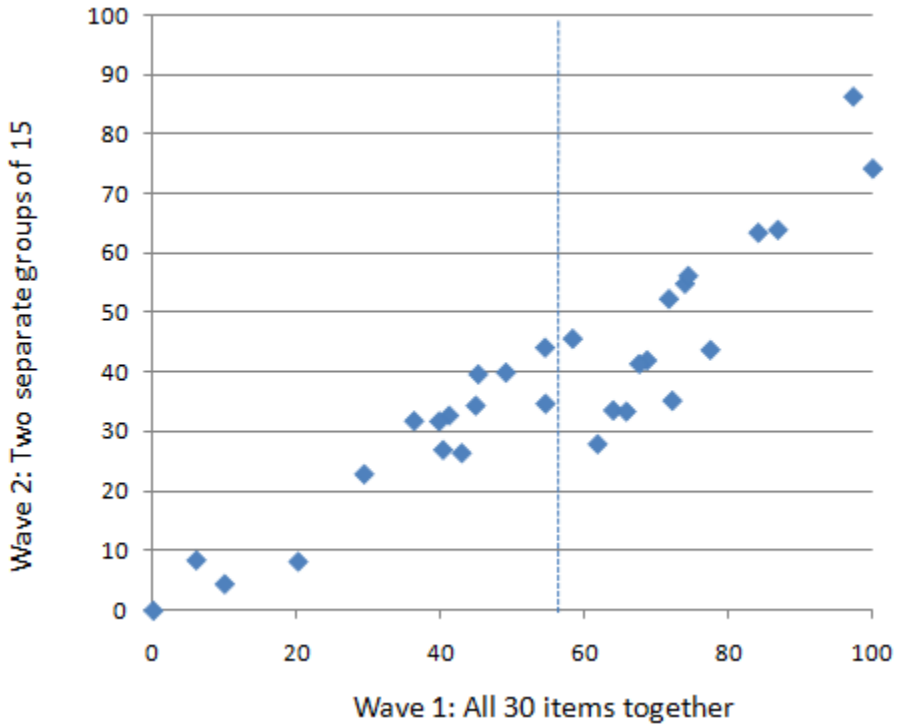
The data were collected using hotspex's Internet Panel. The sample consisted of about 600 Canadian respondents, age 18+. The data collection went very smoothly, with Wave 2 fielded just a few days after Wave 1. Again, we used the utilities from Wave 1 to separate the items into two groups (top 15 and bottom 15 items). Respondents in Wave 2 received Adaptive MaxDiff questionnaires involving either the top15 or the bottom 15 items.

There were a few differences in our execution of Study 2 compared to Study 1. First, we interviewed the same respondents in Wave 2 as Wave 1. We decided only to employ the most effective calibration method from Study 1 (involving the paired comparison judgments of open-end items vs. standard MaxDiff items). Rather than eliciting four open-ended items (two most important and two least important), we decided to elicit just three open-ended items (two most important and one least important). We did this because we noted in Study 1 that when respondents were unable to supply a complete set of legitimate open-end responses, it was typically the worst items they had difficulty specifying (as a result, only 10% of respondents in Study 2 couldn't provide a full set of legitimate answers—a slight improvement over the 15% rate for Study 1). Rather than using a total of eight paired-comparison judgments between open-end items and standard MaxDiff items, we included just six paired-comparison judgments in Study #2 (each of the open-end items versus two of the standard items).

[6] Lynd Bacon provided this example in his critique of our work.

After estimating the utilities using aggregate logit and rescaling the scores to range from 0-100 (then dropping the open-end item scores in Wave 2 results), the rescaled scores are compared in Figure 5:

Figure 5: Study 2, Wave 1 vs. Wave 2 Scores



The R-Squared for the regression line fitting the two waves of data is 0.85. Again, it appears the calibration approach applied to the two halves of data from Wave 2 has done a creditable job capturing the original scaling of all items from Wave 1.

There still appears to remain some vertical shift in scores from the worst 15 items relative to the best 15 items, though it is slight. We might have expected even better fit than with Study 1 (0.91), since Study 2's sample size was nearly triple its size and we interviewed the same respondents in Wave 1 and Wave 2. Perhaps the use of just three open-end items and six paired comparisons led to the slightly lower fit, or maybe there were stronger context effects. Until we see more evidence, we recommend that researchers looking for a remedy to the relative scaling issue in MaxDiff consider using the calibration approach of Study 1: four open-ended items with eight paired comparison calibration questions.

References

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

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.

Orme, Bryan (2006), "Adaptive Maximum Difference Scaling," Sawtooth Software Research Paper, available at www.sawtoothsoftware.com.

Orme, Bryan and Christopher King (2008), "Political Landscape 2008: Segmentation Using MaxDiff and Cluster Ensemble Analysis," *Alert! Magazine*, Marketing Research Association, October 2008, Vol. 46 No.10.

Sawtooth Software (2007), "MaxDiff/Web Technical Paper," available at www.sawtoothsoftware.com.

Appendix A: Study 1 Sample Questions

Example MaxDiff question:

Of these five, pick your most and least preferred for our leaders to accomplish or concentrate on.

Most Preferred	Least Preferred	
○	○	Develop alternative energy sources
○	○	Restrict gun ownership
○	○	Improve infrastructure such as roads and rails
○	○	Reduce taxes for middle and lower income households
○	○	Enact policies to solve housing / mortgage crisis

Example elicitation of open-end responses for top two preferred issues:

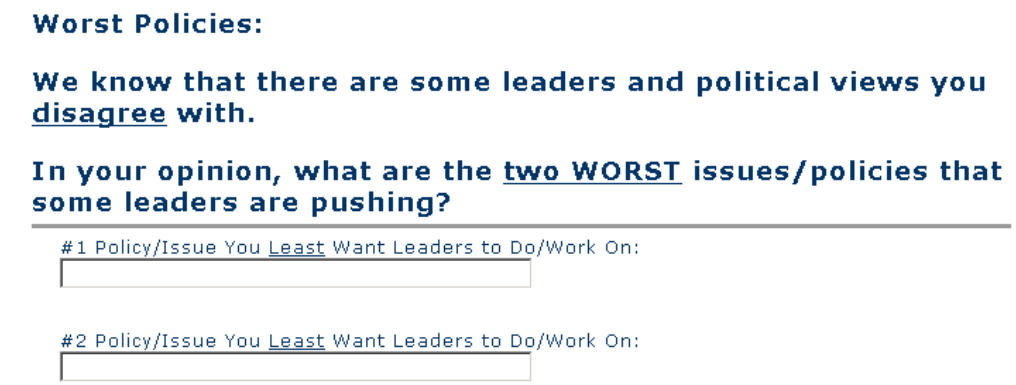
Critical Issues:

In your opinion, what are the two most critical issues/policies that you'd prefer the leaders of our nation accomplish or concentrate on solving?

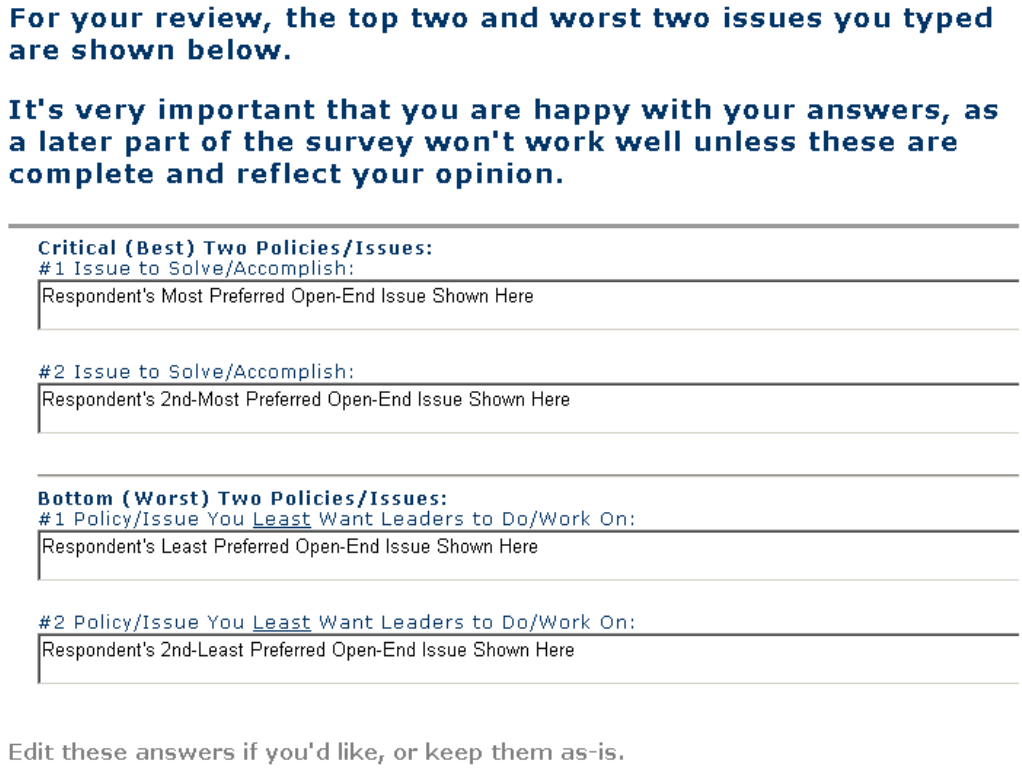
#1 Issue to Solve/Accomplish: _____

#2 Issue to Solve/Accomplish: _____

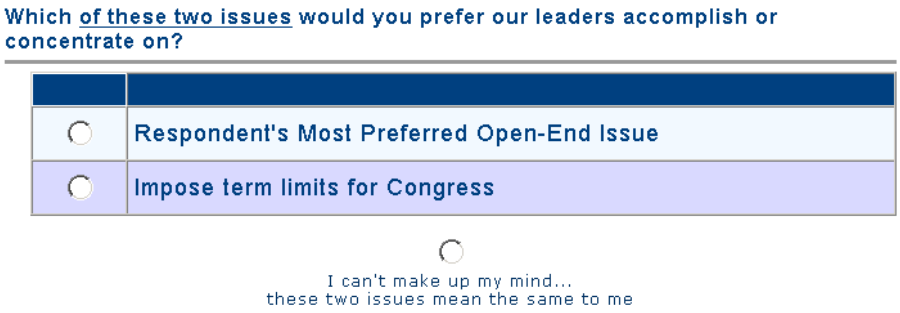
Example elicitation of open-end responses for two least preferred issues:



Example reminder screen that allows respondents to review their open-end answers and make any needed changes (we showed this to wave 2 respondents after three MaxDiff tasks had exposed respondents to all 15 items used in their questionnaire):



Example Calibration Screen for Method of Paired Comparisons (MPC), comparing respondent's open-end extreme items to items included in the MaxDiff experiment:



Note that whatever the respondent had typed earlier in the questionnaire as their Most Preferred item is dynamically inserted in place of "Respondent's Most Preferred Open-End Issue" (as shown here, for illustration). So, if a respondent says "We need to fix this broken economy!" then that exact text is displayed as the first item within this pair.

Appendix B: Adaptive MaxDiff Design

Round I (Best/Worst)		Round II (Best/Worst)		Round III (Best/Worst)		Round IV (MPC)		Round V (Best/Worst)	
Set 1	Item 1 Item 2 Item 3 Item 4 Item 5	Set 4	Set 1 winner Set 2 item Set 3 item Set 1 item	Set 7	Set 4 winner Set 5 item Set 6 item	Set 10	Set 7 winner Set 8 item	Set 13	Set 10 winner Set 11 winner Set 12 winner
Set 2	Item 6 Item 7 Item 8 Item 9 Item 10	Set 5	Set 2 winner Set 3 item Set 1 item Set 2 item	Set 8	Set 5 winner Set 6 item Set 4 item	Set 11	Set8 winner Set 9 item		
Set 3	Item 11 Item 12 Item 13 Item 14 Item 15	Set 6	Set 3 winner Set 1 item Set 2 item Set 3 item	Set 9	Set 6 winner Set 4 item Set 5 item	Set 12	Set 9 winner Set 7 item		

As an example, Set 4 includes the winning item from Set 1 (the item chosen as most preferred from Set 1), plus a surviving item (an item not chosen as either most or least preferred in a previous set) drawn randomly (without replacement) from sets 2, 3, and 1. Items 1 through 15 are initially randomized before placing them in Round I. And, of course, the items within each subsequent set are randomized so that the winning items are not always displayed in first position as shown in the grid above.

Appendix C: Study 2 Example MaxDiff Question

www.hotspex.com

Of these five factors, which is the most important in affecting your confidence as a consumer and which is the least important?

Most Important		Least Important
○	There is ready access to capital / funding for business development	○
○	You have ample retirement savings	○
○	You have a low balance owing on your credit card(s)	○
○	Your current company has a healthy and stable place in the market	○
○	There are effective plans in place to take care of the elderly / retiring baby boomers	○