Sawtooth Software

RESEARCH PAPER SERIES

Adaptive Maximum Difference Scaling

Bryan Orme, Sawtooth Software

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

Adaptive Maximum Difference Scaling

Bryan Orme, President
Sawtooth Software

May, 2006

Background

One of the most common tasks for researchers is to measure the importance or desirability of a list of items. A relatively new methodology for accomplishing this called Maximum Difference Scaling (also known as MaxDiff or best-worst scaling) is gaining momentum in our industry. In fact, papers on MaxDiff have won the "best presentation" award at recent ESOMAR and Sawtooth Software conferences (Cohen and Markowitz 2002, Cohen 2003, and Chrzan 2004). A recent study by Cohen and Orme showed that MaxDiff performed better than standard rating scales in terms of discriminating among items, discriminating between respondents on the items, and predictive validity of holdout (ranking) questions (Cohen and Orme 2004).

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).

An example MaxDiff question (from the current research) is shown below:

Exhibit 1

When considering a laptop computer to buy,
among the five features shown here,
which are the most and least important to you?

Most Important		Least Important
○	Lots of memory (RAM)	○
○	Larger than typical hard drive	○
○	Video card configured for optimal gaming performance	○
○	Includes a DVD burner	○
○	Fast processor speed (MHz rating)	○

MaxDiff tasks (sets) such as these are repeated, until each respondent typically sees each item at least 3 times (in the case that individual-level estimation of item scores is desired). With a 20-item study displaying 5 items per set, this would involve at least 12 sets. The sets are designed with the goal that each item appear an equal number of times and that each item appear with every other item an equal number of times. In addition, designs are chosen such that each item appears in each position in the set an equal number of times. Designs that are not perfectly balanced in all these respects, such as those including item prohibitions, still achieve very respectable precision of estimates (Sawtooth Software 2005).

MaxDiff has also been proposed for traditional conjoint analysis problems (where the level weights are summed across multiple attributes to predict preference for a product whole). However, the MaxDiff question device and model specification do not formally consider or justify additivity of the levels. Although some researchers have advocated MaxDiff in conjoint-style applications (so-called "best-worst conjoint"), we do not suggest MaxDiff in these situations. MaxDiff is appropriate for research problems focusing on the contrasts between typically a dozen or more items in terms of preference or importance, rather than the additive effect across multiple items.

Exhibit 2 summarizes the usual focus and strategies of traditional conjoint/choice models vs. MaxDiff.

Exhibit 2

Conjoint/Choice Models	MaxDiff
Typically 2 to 8 attributes with 2 to 5 levels each	Typically 10 to 30 items
Precise attributes, each with at least 2 levels	Both precise and general "single statement" attributes
Product design, pricing, segmentation	Promotion, image, segmentation
Need to simulate choices among multi-attribute products	Need to measure attention captured by each item
Competition oriented	Orientation to single product

The Case for Adaptive Maximum Difference Scaling (A-MaxDiff)

MaxDiff exercises result in better precision and discrimination of item scores than the typical rating scales used in questionnaires, but they also require more effort and time on the part of the respondent (Cohen and Orme 2004). MaxDiff tasks are very repetitive as well, as each set has the same layout (though each set reflects a new combination of items). Furthermore, the MaxDiff questionnaire places equal attention on estimating the value of worst items as well as best items. It would seem that researchers should be more interested in obtaining the highest precision of estimates for items of greatest value to respondents. Managerial actions usually focus on these. This is especially the case when results from MaxDiff are used in simulation procedures such as TURF, where the algorithm seeks to identify the top few items that can reach and satisfy respondents. For

many applications, perhaps MaxDiff could be improved if it could be made to focus on stabilizing scores for the most important (preferred) items.

In the 1980s, Rich Johnson (founder of Sawtooth Software) developed a popular program for Adaptive Conjoint Analysis (ACA). It became very successful because the interview learned from respondents' previous answers to customize more relevant and efficient tradeoffs in later conjoint questions. In contrast to the standard method of conjoint at the time (full-profile ratings-based conjoint) that repeated the same-looking layout over and over, ACA provided a more engaging experience for respondents. The conjoint portion of the interview (the pairs) typically started with just two attributes at a time, and then gradually included more attributes in later stages. And, later pairs questions increased in relevance and difficulty for the respondent, as the product concepts to be compared became more similar in terms of utility. An experiment published by Huber *et al.* showed that respondents found the adaptive interview to be more enjoyable and it appeared to respondents to take less time than the traditional full-profile conjoint method (Huber *et al.* 1991).

For this current research, we developed a simple adaptive approach for MaxDiff questionnaires (A-MaxDiff). Like ACA, it learns from respondents' previous answers to develop more challenging and utility-balanced tasks in later stages. Utility balance has been shown to be an important element in increasing the efficiency of choice-based questionnaires that employ logit to estimate parameters (Huber and Zwerina 1996). The A-MaxDiff approach also varies the layout of the MaxDiff questions in stages. But, rather than progressively increase the number of attributes in the task (as with ACA), it gradually *decreases* the number of items considered in each task as the level of difficulty (due to utility balance) increases. Furthermore, the amount of information collected (and therefore the efficiency of estimation) is much greater for items of greater importance to respondents, as these are included relatively more times across the sets.

We conducted two field experiments to test the A-MaxDiff methodology. The first we describe in detail below. The second (the findings essentially mirror those of the first) is summarized in Appendix C.

Description of the Experiment

605 respondents were interviewed using GMI's internet panel (Global Market Insite, www.gmi-mr.com) during January, 2006. (We're indebted to GMI for supporting this methodological research.) The questionnaire was programmed using Sawtooth Software's SSI Web system (v5.4). All aspects of the A-MaxDiff questionnaire were programmed using the sophisticated list-building logic (piping) available within the CiW component of SSI Web. No customization of the SSI Web system was required to implement the essentials of the A-MaxDiff questionnaire.

Respondents were randomly divided into two groups: one receiving standard MaxDiff (5 items per set for 16 sets) and the other receiving A-MaxDiff (varying between 2 and 5 items per set). The questionnaires were (virtually) length-equalized in terms of the

number of clicks required to complete the MaxDiff exercises (32 clicks for standard MaxDiff and 31 clicks for A-MaxDiff). Both groups of respondents received two holdout questions (choices of best and worst from seven items). The holdout tasks occurred after the fourth MaxDiff question had been asked.

The sample sizes for the two groups of respondents were:

A-MaxDiff	303
MaxDiff	302

The experiment involved obtaining importance scores for 20 items related to purchasing laptop computers (to be computed using HB analysis). The primary goal was to obtain the highest precision for items of greatest importance to each respondent, and the secondary goal was to obtain robust scores, at least at the group level, for all 20 items in the experiment, as shown below:

Exhibit 3

Items in MaxDiff Experiment

1. Fast processor speed (MHz rating)
2. Lots of memory (RAM)
3. Superior technical support
4. Video card configured for optimal gaming performance
5. Larger than typical hard drive
6. Ability to purchase in a store near you
7. Ability to purchase over the internet
8. Comes with industry standard anti-virus/anti-spyware software
9. Includes Microsoft Office
10. Light weight model
11. Integrated subwoofer for improved sound
12. Double the normal warranty
13. Includes a DVD burner
14. Includes Windows XP Professional (rather than Home Edition)
15. Includes TV tuner (receives antenna or cable input)
16. Has an additional battery
17. Comes with 6 months America Online access
18. Comes with $250 rebate
19. Laser cordless mouse included
20. Mfr offers 8% financing for 48 months

The A-MaxDiff Questionnaire

We employed a five-stage adaptive MaxDiff (A-MaxDiff) questionnaire. It was adaptive in the sense that as the questionnaire progressed, items of lesser importance were discarded from further consideration. Toward the end of the A-MaxDiff questionnaire,

respondents compared the few surviving items of greatest importance to them. The later questions had the benefit of increased utility balance, which has a statistical benefit to efficiency for choice-based models (Huber and Zwerina 1996). Of course, as respondents are asked to compare items that are all relatively important, the questions become more difficult. To counteract that difficulty, the number of items presented in each set (task) was progressively reduced from the starting five to eventually just two per set. This allows respondents to concentrate on discriminating among nearly equally important items in a more focused setting involving fewer items (only paired comparisons in the final stage). The amount of information for items of greater importance is also increased (relative to standard MaxDiff) since these items occur more throughout the interview.

How does the A-MaxDiff questionnaire adapt to the respondent's preference and discard less-important items from further consideration? We began showing items five-at-a-time in each set, which has been shown to be nearly optimal in previous research (Orme 2005), (Chrzan and Patterson 2006). Whenever an item was chosen as "worst" in a best/worst task, it was deleted from further consideration. The questionnaire had different stages, wherein the number of items shown per set was reduced by one in each stage.

With 20 items in this exercise, each item initially can be displayed one time using 4 sets with 5 items in each set. That comprised Stage I. Stage II only involved the 16 surviving items (those marked least important in Stage I were deleted from further consideration). A summary of the complexity of the sets in each stage is shown in Exhibit 4.

Exhibit 4

Stage	Set Composition within Stage	Surviving Items in Stage
Stage I	4 MaxDiff sets with 5 items per set	20
Stage II	4 MaxDiff sets with 4 items per set	16
Stage III	4 MaxDiff sets with 3 items per set	12
Stage IV	4 Pairs with 2 items per pair	8
Stage V	Final ranking of surviving 4 items	4

The grid below describes the adaptive interview design in more detail. Item order is randomized for each respondent going into Stage I, with the remaining design following an experimental plan as shown in Exhibit 5 below.

Exhibit 5

Stage I (Best/Worst)		Stage II (Best/Worst)		Stage III (Best/Worst)		Stage IV (MPC)		Stage V (Rank)	
Set 1	Item 1 Item 2 Item 3 Item 4 Item 5	Set 5	Set 1 winner Set 2 item Set 3 item Set 4 item	Set 9	Set 5 winner Set 6 item Set 7 item	Set 13	Set 9 winner Set 10 item	Set 17	Set 13 winner Set 14 winner Set 15 winner Set 16 winner
Set 2	Item 6 Item 7 Item 8 Item 9 Item 10	Set 6	Set 2 winner Set 1 item Set 3 item Set 4 item	Set 10	Set 6 winner Set 7 item Set 8 item	Set 14	Set 10 winner Set 11 item	Set 18	Set 17 item Set 17 item
Set 3	Item 11 Item 12 Item 13 Item 14 Item 15	Set 7	Set 3 winner Set 1 item Set 2 item Set 4 item	Set 11	Set 7 winner Set 8 item Set 5 item	Set 15	Set 11 winner Set 12 item		
Set 4	Item 16 Item 17 Item 18 Item 19 Item 20	Set 8	Set 4 winner Set 1 item Set 2 item Set 3 item	Set 12	Set 8 winner Set 5 item Set 6 item	Set 16	Set 12 winner Set 9 item		

As an example, Set 5 includes the winning item from Set 1 (the item chosen as most important from Set 1), plus a surviving item (an item not chosen as either most or least important in a previous set) drawn randomly (without replacement) each from sets 2, 3, and 4. And, of course, the items within each set are randomized so that the winning items are not always displayed in first position as shown in the grid above.

For the design used in this experiment, the number of times each item occurred varied greatly at the individual level. Those items marked least important in the first stage only were shown once to the respondent; whereas, items surviving to the final stage were shown either 5 or 6 times. (For the standard MaxDiff experiment, each item occurred exactly 4 times for each respondent.) If the fact that some items were only shown once concerns you, you could add another stage (Stage 0) at the beginning of the experiment that also showed 5 items per set. Then, 4 items could be eliminated from further consideration after all items had been shown twice (after stages 0 and I). The remainder of the experiment would continue as shown in Exhibit 5.

It is important to note that the design as described here doesn't *guarantee* that the top 4 items in importance for each respondent are compared in Stage V. Assuming the respondent answers in a consistent manner, it is possible for a third- or fourth-place item not to make it into Stage V. However, this design ensures that the top two items in terms of importance are preserved to the final ranking in the final stage. Because score estimation is performed based on the logit rule (over all sets), a third- or fourth-place item not surviving to the ranking exercise in the final stage can still receive a third- or fourth-place score in terms of the final logit estimation. Using a sports team ranking analogy, a third-place item that gets eliminated due to being compared with the first- or second-

place item in Stage IV (the only stage in which such an elimination could occur in this design) gets credited with a "quality" loss.

Between stages of the questionnaire, we included some text to encourage respondents and prepare them for the next section. We included similarly appropriate statements and placed them in the same relative positions within the standard MaxDiff questionnaire. These prompts are described in Appendix A.

Test Results

Using HB, we estimated (using Sawtooth Software's CBC/HB program) parameter scores for the 20 items separately for respondents receiving standard MaxDiff or A-MaxDiff. A description of the appropriate design matrix coding to use within HB estimation for MaxDiff and paired comparison tasks is given in the "MaxDiff/Web System Technical Paper," available at www.sawtoothsoftware.com (Sawtooth Software 2005). To eliminate the possibility that differences between respondents may be confounded by scale, we normalized the parameter estimates for each respondent. The normalization procedure involved subtracting off a constant for all 20 parameters such that the least important score was set to zero, followed by multiplying all 20 parameters by a constant such that the most important score was 100. This is identical to the method of normalization called "Points" used in older versions of Sawtooth Software's market simulator for conjoint data.

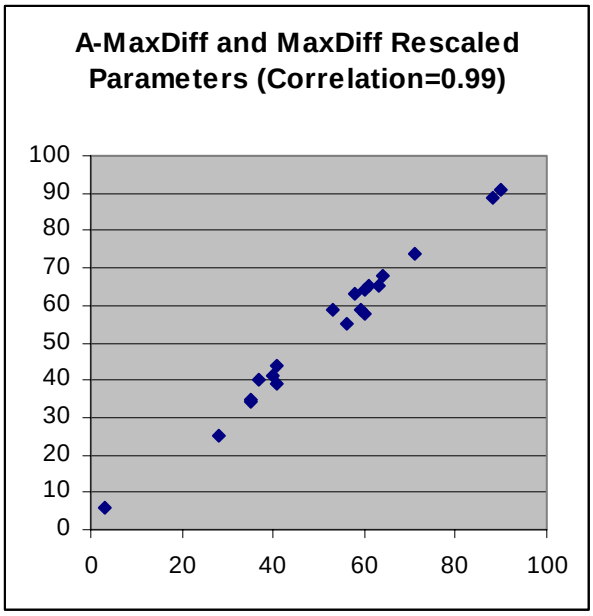
The mean parameters were nearly identical, as can be seen in Exhibits 6 and 7.

Exhibit 6

Rescaled HB Parameters
(Normalized to Range 0-100 for Each Respondent)

	A-MaxDiff	MaxDiff
Fast processor speed (MHz rating)	90	91
Lots of memory (RAM)	88	89
Superior technical support	60	58
Video card configured for optimal gaming performance	41	44
Larger than typical hard drive	71	74
Ability to purchase in a store near you	40	41
Ability to purchase over the internet	41	39
Comes with industry standard anti-virus/anti-spyware software	56	55
Includes Microsoft Office	58	63
Light weight model	59	59
Integrated subwoofer for improved sound	35	35
Double the normal warranty	63	65
Includes a DVD burner	61	65
Includes Windows XP Professional (rather than Home Edition)	60	64
Includes TV tuner (receives antenna or cable input)	35	34
Has an additional battery	53	59
Comes with 6 months America Online access	3	6
Comes with $250 rebate	64	68
Laser cordless mouse included	37	40
Mfr offers 8% financing for 48 months	28	25

Exhibit 7



Hit rates are a commonly used measure to gauge the predictive accuracy of the estimated scores to predict some additional question or outcome not used (held out) for the estimation of the scores. Recall that 2 holdout sets of 7 items were included in the

questionnaire. The items in these sets were drawn randomly for each respondent. Respondents were asked to indicate the most and least important item in each holdout set. If the estimated scores predict the respondent's answer correctly, then we score a "hit," otherwise we score a "miss." The hit rate is the percent of choices predicted correctly.

Hit rates were as follows, with standard errors shown in parentheses:

Exhibit 8

	A-MaxDiff	MaxDiff
Choice of Bests*	70.5% (1.94)	64.2% (2.09)
Choice of Worsts**	59.6% (2.06)	66.6% (2.00)

* Significant difference between bests at $t = 2.21$ ($p<0.05$)
** Significant difference between worsts at $t = -2.44$ ($p<0.05$)

These hit rates are appreciably better than that expected due to chance (1/7=14%). The A-MaxDiff method results in more accurate identification of most important items, but less accurate identification of least important items. We also examined the likelihood of respondents having chosen the item they selected as best (according to the logit rule) and found that the difference in likelihoods between A-MaxDiff and standard MaxDiff was significant at $p<0.01$. Of course, if we had been primarily interested in the precision of estimates for least important (least desirable) items, we could have flipped the questioning around, eliminating items of greatest importance from further consideration.

To ensure that the results seen above were not due to an unlucky assignment of more consistent respondents into one group over the other, we referred to a separate test-retest reliability estimate within the same questionnaire. The second half of the questionnaire actually involved a separate CBC experiment not reported here. Within that experiment, we repeated two holdout CBC questions involving similar attributes related to laptop purchases. The test-retest reliability for the two groups of respondents reported here on those repeated CBC questions was 72.1% and 72.2%, respectively. Although the CBC tasks involved a different style question, it seems to be a reasonable proxy to demonstrate that both groups of people exhibited equal tendency to be internally consistent in their answers to tradeoff questions involving laptops.

Recall that the experiment was designed to equalize the number of clicks required to complete both A-MaxDiff and standard MaxDiff exercises. To answer a MaxDiff question requires two clicks (one click to indicate the most important item, and one to indicate the least important item). Questions late in the A-MaxDiff questionnaire involved choices from pairs of items (2 items at a time), which of course involve just one click.

With help from my programming colleagues here at Sawtooth Software, we captured the time it took for respondents to submit each question in the survey (using Javascript). We only counted the time from when the page was loaded to the time the respondent clicked

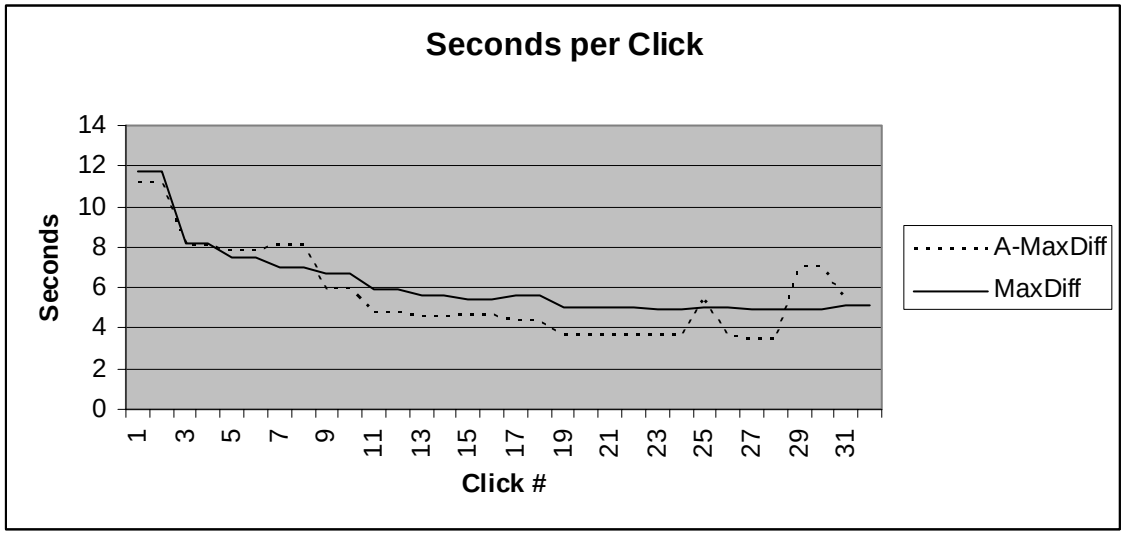
the Next button. We did not include any transmission time to and from the server, but testing during the highest traffic period during data collection suggested the transmission was nearly instantaneous over a high-speed connection. The median time required per click is reported in Exhibit 9, along with the task complexity (number of items being evaluated for each click). Note that for MaxDiff tasks, one time measurement was captured per page submission, and therefore we simply split the time equally across clicks representing most important and least important items for that page.

Exhibit 9
Interview Duration
Seconds per Click

Click#	Time (seconds) A-MaxDiff	# Items/ Set	Time (seconds) MaxDiff	# Items/ Set
1	11.1	5	11.7	5
2	11.1	5	11.7	5
3	8.1	5	8.2	5
4	8.1	5	8.2	5
5	7.8	5	7.5	5
6	7.8	5	7.5	5
7	8.1	5	7.0	5
8	8.1	5	7.0	5
9	5.9	4	6.7	5
10	5.9	4	6.7	5
11	4.7	4	5.9	5
12	4.7	4	5.9	5
13	4.5	4	5.6	5
14	4.5	4	5.6	5
15	4.6	4	5.4	5
16	4.6	4	5.4	5
17	4.3	3	5.6	5
18	4.3	3	5.6	5
19	3.6	3	5.0	5
20	3.6	3	5.0	5
21	3.6	3	5.0	5
22	3.6	3	5.0	5
23	3.6	3	4.9	5
24	3.6	3	4.9	5
25	5.3	2	5.0	5
26	3.6	2	5.0	5
27	3.5	2	4.9	5
28	3.5	2	4.9	5
29	7.0	4	4.9	5
30	7.0	4	4.9	5
31	5.3	2	5.1	5
32			5.1	5
Total:	175.0 (2.92 minutes)		196.8 (3.28 minutes)	

Since this table reflects a great deal of detail, we've charted the results in Exhibit 10.

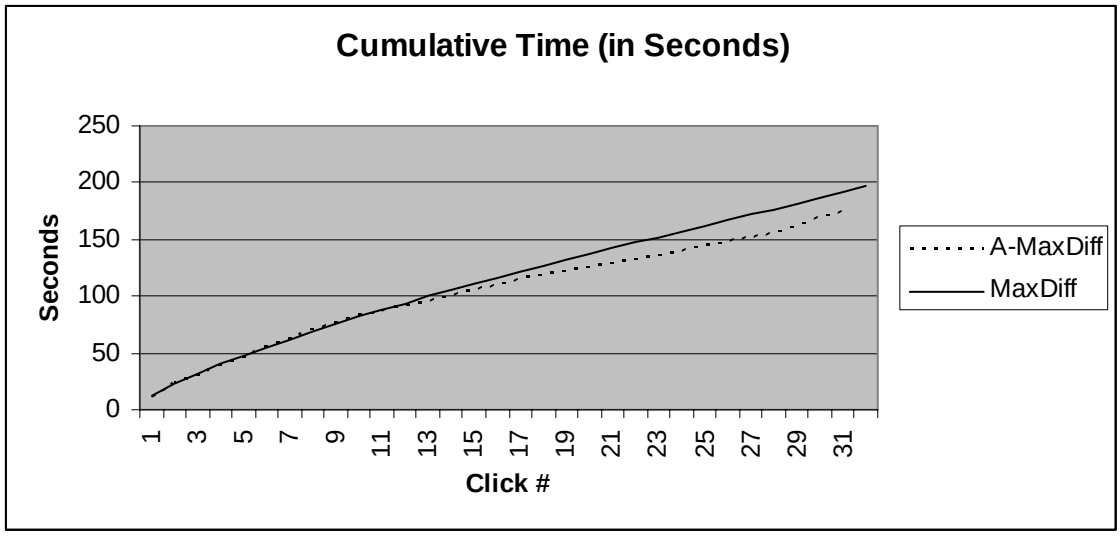
Exhibit 10



For the 1st through 8th clicks, both MaxDiff methods showed 5 items per set. On the 9th through 16th clicks, A-MaxDiff presents 4 items at a time, and we begin to see a time savings compared to the standard MaxDiff exercise. On the 18th through 24th clicks, A-MaxDiff presents 3 items at a time. The time per click spikes slightly at the 25th click for A-MaxDiff, because respondents need to orient themselves to a new format displaying 2 items at a time (horizontal format rather than vertical). Time spikes again for A-MaxDiff at the 29th and 30th clicks, which change back to 4 items at a time. On the 31st click, respondents see 2 items at a time again.

The cumulative time for both flavors of MaxDiff questionnaires is shown in Exhibit 11.

Exhibit 11



The A-MaxDiff exercise took 175 seconds in total (2.92 minutes), whereas the MaxDiff exercise took 196.8 seconds (3.28 minutes).

In addition to the quantitative measures of performance for A-MaxDiff, we also asked respondents their perceptions of each task. We borrowed this qualitative assessment from the 1991 paper by Huber *et al.*, who asked the same questions of respondents completing ACA vs. traditional full-profile conjoint.

Exhibit 12
Qualitative Assessment

Using a scale where a 1 means "strongly disagree" and a 7 means "strongly agree," how much do you agree or disagree that the previous series of questions you just answered (where we showed you features and you selected most/least important)...		
	A-MaxDiff	**MaxDiff**
…was enjoyable*	5.53 (0.078)	5.28 (0.082)
…was confusing	1.96 (0.091)	1.87 (0.082)
…was easy	5.96 (0.076)	5.96 (0.071)
…made me feel like clicking answers just to get done*	2.25 (0.097)	2.47 (0.010)
…allowed me to express my opinions	5.65 (0.082)	5.55 (0.084)

* indicates significant difference at $p<0.05$.

Both versions of MaxDiff seemed enjoyable and easy to respondents. Respondents didn't rate either task as very confusing. There are two statistically significant differences between the MaxDiff methods. Respondents perceived A-MaxDiff to be a bit more enjoyable and a bit less monotonous than regular MaxDiff, for which respondents were more likely to express that they felt like "clicking answers just to get done." With the increased utility balance as respondents went deeper into the A-MaxDiff questionnaire, one would think that respondents would have perceived the interview to be more difficult. That would not appear to be the case, as respondents found either exercise equally easy (5.96 on a 7-pt scale). The shorter length of the A-MaxDiff questionnaire and the lower complexity (showing fewer items at a time) seemed to counteract the increased difficulty due to the utility balance.

Summary

This research paints a favorable picture for Adaptive MaxDiff questionnaires. Relative to a best-practices implementation of the standard MaxDiff approach, A-MaxDiff:

- Leads to very similar mean parameter estimates
- Takes less time to complete the same number of questions (clicks)
- Yields more precise identification of each respondent's most important items
- Is perceived to be more enjoyable and less monotonous

These findings were confirmed by a second field test reported in Appendix C.

The drawbacks for A-MaxDiff are:

- It leads to less precise identification of each respondent's least important items
- It is more complicated to program than standard MaxDiff exercises

Any researcher using a capable CAPI or Web-interviewing solution that supports list-building logic (piping) can implement A-MaxDiff. Instructions for generalizing the experiment to numbers of items different from 20 are provided in Appendix B.

References

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

Chrzan, Keith and Michael Patterson (2006), "Testing for the Optimal Number of Attributes in MaxDiff Questions," *2006 Sawtooth Software Conference Proceedings*, Forthcoming.

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.

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.

Huber, Joel C., Dick Wittink, John Fiedler, and Richard Miller (1991), "An Empirical Comparison of ACA and Full Profile Judgments," 1991 Sawtooth Software Conference Proceedings, 189-202.

Huber, Loel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, (August), 303-317.

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.

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

Sawtooth Software (2005), "The MaxDiff/Web System Technical Paper," Technical Paper available at www.sawtoothsoftware.com.

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

Appendix A
Prompts Used within MaxDiff Questionnaire

Prior to the first question:

Both A-MaxDiff and MaxDiff:

> We want to learn what aspects are important to you when purchasing a laptop (notebook) computer. To do this, we are going to ask you a series of tradeoff questions. We'll show you some sets of features and ask which are the most and least important to you. We need to ask you repeated questions to learn how you make these sometimes complex tradeoffs.

After first four MaxDiff questions have been completed:

A-MaxDiff:

> Now that we've seen a few of your answers, we are starting to learn what you think is important. This time, we'll just show four items at a time.

Standard MaxDiff:

> Now that we've seen a few of your answers, we are starting to learn what you think is important. You've completed four of these kinds of questions, and we need to ask you some more to learn even more about what features are most and least important to you. Keep up the good work!

After eight MaxDiff questions have been completed:

A-MaxDiff:

> OK, now we are becoming more certain regarding what is important to you. The next questions will display just three items at a time. We've discarded the features you feel are less important, and we'll just be asking you to trade off the more important items.

Standard MaxDiff:

> OK, now we are becoming more certain regarding what is important to you. Just eight more questions like this in this section. Please concentrate and answer as best you can.

After twelve MaxDiff questions have been completed:

A-MaxDiff:

> You are just about done with this section. The last few questions focus on the most important features to you. Hang in there!

And again, right prior to the last two questions:

We're really narrowing it down now. Just two more questions in this section. These last two questions are very important to us, because they tell us what feature is most important to you!

Standard MaxDiff:

You are just about done with this section. Just four more questions in this section to go. Hang in there!

Appendix B
Generalizing the Adaptive MaxDiff Design Method

One can generalize an A-MaxDiff plan to accommodate any number of items divisible by 4, 5, or 6. It makes most sense to use numbers of items divisible by 4, 5, or 6 because previous research has suggested that MaxDiff experiments are most efficient when they include from 4 to 6 items per set. To design an A-MaxDiff plan, one must specify:

t = Total number of items
i = Maximum number of items to show per set (must evenly divide into t and be equal to 4, 5, or 6)
c = Minimum number of times each item should appear in the experiment for each respondent (suggested number is 1 or 2)

With that information, the first stage is designed as follows:

The number of sets, s_1, in Stage I should be equal to tc/i. Randomize the list of items and arrange these items into s_1 sets of size i. In building sets, once all items have been used once, then (if c>1) randomize the full list again and continue random assignment.

The least important t/i items are identified based on how many times they were marked as "worst" in Stage I (ties are broken randomly in the case of c>1). These t/i items are not carried forward into subsequent stages.

In all subsequent stages, each item appears only once within sets for that stage.

The number of sets, s_2, for stage II is equal to the number of sets from Stage I, except that each set now includes i-1 items. Assign a winning item from the previous stage to each of the sets (one winner per set). For example, the winner from Set 1 of Stage I is assigned to Set 1 of Stage II, etc. Assign other surviving items to the sets as follows: Set1 contains one item (not previously selected into any Stage II set) drawn randomly from Set 2 of the previous stage, one item drawn randomly from Set 3 of the previous stage, etc. until each set in Stage II contains (i-1) items. Set 2 contains one item (not previously selected into any Stage II set) drawn randomly from Set 1, one item drawn randomly from Set 3, etc. (wrapping around to Set 1 once the last set has been referenced). Follow this pattern to assign all items to s_2 sets for Stage II.

Repeat a similar procedure for subsequent stages. When sets include just 2 items, only "best" questions (paired comparisons) are asked (it is redundant to ask for "worsts"). Non-selected items from pairs are not carried forward to subsequent stages.

Once the stage displaying 2 items per set (pairs) has been asked, we cut to the chase to identify the winning item:

#Surviving Items	Strategy
2	Ask 1 pair with 2 items
3	Ask 1 best-worst question with 3 items
4	Ask 1 best-worst question with 4 items and (using surviving items) 1 pairs question
5	Ask 1 best-worst question with 3 items, 1 pairs question, and (using surviving items) 1 best-worst question with 3 items
6	Ask 3 pairs questions with 2 items each, and (using surviving items) 1 best-worst question with three items
7	Ask 1 best-worst question with 3 items, 2 pairs questions, and (with surviving items), 1 best-worst question with 4 items and (using surviving items) 1 pairs question
8	Ask 4 pairs questions, and (with surviving items) 1 best-worst question with 4 items and (using surviving items) 1 pairs question
9	Ask 3 best-worst questions with three items, and (with surviving items) 3 pairs questions, and (with surviving items) ask 1 best-worst question with 3 items
10	Ask 5 pairs questions, and (using surviving items) ask 1 best-worst question with 3 items, 1 pairs question, and (using surviving items) 1 best-worst question with 3 items

It's unusual to have more than 10 surviving items from the stage with paired comparisons. However, one can see the patterns to follow (to narrow down the one winner) should that occur.

Number of Items in Symmetric A-MaxDiff Exercises

So that the Adaptive MaxDiff exercise is symmetric (each surviving item shown an equal number of times within each stage), only certain numbers of total items are supported in the exercise.

# Items in Exercise	Suggested Beginning Set Size
4	4
5	5
6	6
8	4
10	5
12	6 preferred, 4 also possible
15	5
16	4
18	6
20	5 preferred, 4 also possible
24	6 preferred, 4 also possible
25	5
28	4

30	6 or 5
32	4
35	5
36	6 preferred, 4 also possible
40	5 preferred, 4 also possible
42	6
44	4
45	5
48	6 preferred, 4 also possible
50	5

Adaptive MaxDiff surveys are not very feasible for numbers of items beyond 50. A 50-item A-MaxDiff study would involve 80 clicks (responses). This probably exceeds what is reasonable to ask respondents to complete.

Non-symmetric A-MaxDiff designs should also be possible. The benefits of such plans would be no restrictions in choosing the number of items in the study and also flexibility in choosing the beginning set size. The main drawback would be greater complexity of programming.

Appendix C
Second MaxDiff Test (Fast-Food Restaurant Features)

In May, 2006, we conducted a second test of Adaptive MaxDiff, this time focusing on eighteen fast-food restaurant features. We appreciate Western Watts for donating the sample from their Opinion Outpost Global Survey Research Panel.

As with the first test, we equalized the number of clicks between the adaptive and standard MaxDiff questionnaires (28 clicks for MaxDiff, 29 for A-MaxDiff). The MaxDiff exercise showed 5 items per set, for 14 sets (each item shown an average of 3.9 times). The A-MaxDiff exercise started with 6 items per set in the first stage. As with the first test, we employed two holdout sets of seven items.

The research conclusions are essentially identical to the first test:

- The parameters were very similar for adaptive and standard MaxDiff.
- Respondents found the adaptive interview more enjoyable and less monotonous. Additional differences were detected that were not found in the first test: the adaptive interview was seen as easier and allowing respondents to better express their opinions.
- The hit rates for predicting best items from holdouts were significantly higher for A-MaxDiff.

A summary of the design and significant findings is given below:

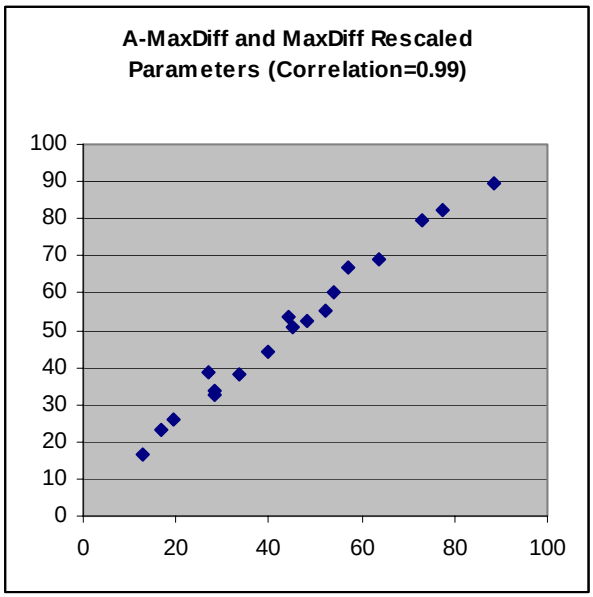
Stage	Set Composition within Stage	Surviving Items in Stage
Stage I	3 MaxDiff sets with 6 items per set	18
Stage II	3 MaxDiff sets with 5 items per set	15
Stage III	3 MaxDiff sets with 4 items per set	12
Stage IV	3 MaxDiff sets with 3 items per set	9
Stage V	3 Pairs with 2 items per pair	6
Stage VI	Final ranking of surviving 3 items	3

Sample Sizes:

A-MaxDiff	331
MaxDiff	350

Rescaled HB Parameters
(Normalized to Range 0-100 for Each Respondent)

	A-MaxDiff	MaxDiff
Great tasting food	89	89
Great value for the money	78	82
Offers program for receiving discounts after a certain number of purchases	27	39
Many healthy options on the menu	54	60
Restaurant gives generously to charities	19	26
Restaurant leads industry in environmental responsibility	29	34
Restaurant known as good employer with excellent employee relations	33	38
Child-friendly atmosphere	28	33
Play area for children	13	17
Fast service (order and get food in 5 minutes or less)	57	67
Employees communicate clearly in your native language	52	55
Workers always fill orders correctly	73	80
Your order is brought to your table	17	23
Cleaner than average eating areas	64	69
Cleaner than average restrooms	48	52
Has a drive-through	45	51
Deep-fries using low trans-fat oil	40	44
Flame-broils meat (rather than fried)	44	54



Hit Rates

	A-MaxDiff	MaxDiff
Choice of Bests*	71.1% (1.78)	63.3% (1.87)
Choice of Worsts**	55.0% (1.88)	61.6% (1.78)

* Significant difference between bests at t = 3.02 ($p<0.01$)
** Significant difference between worsts at t = -2.55 ($p<0.05$)

Qualitative Assessment

Using a scale where a 1 means "strongly disagree" and a 7 means "strongly agree," how much do you agree or disagree that the previous series of questions you just answered (where we showed you features and you selected most/least important)...		
	A-MaxDiff	**MaxDiff**
…was enjoyable*	5.09 (0.075)	4.75 (0.087)
…was confusing	2.05 (0.084)	2.24 (0.085)
…was easy*	5.91 (0.070)	5.69 (0.080)
…made me feel like clicking answers just to get done*	2.18 (0.081)	2.45 (0.089)
…allowed me to express my opinions*	5.53 (0.080)	5.24 (0.084)

* indicates significant difference at $p<0.05$.

Median Time to Complete Exercise

A-MaxDiff	220.1 seconds
MaxDiff	224.0 seconds

(Note: the time savings for A-MaxDiff was not as great as seen in the first test. This is not surprising given that A-MaxDiff started with 6 items per task in the first stage, rather than with 5 items per task in the first stage as in the first test.)