Sawtooth Software

RESEARCH PAPER SERIES

Testing for the Optimal Number of Attributes in MaxDiff Questions

Keith Chrzan, Maritz Research
Michael Patterson, Probit Research

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

Testing for the Optimal Number of Attributes in MaxDiff Questions

Keith Chrzan, Maritz Research
Michael Patterson, Probit Research
March 2006

Introduction

Maximum difference ("maxdiff") scaling (Finn and Louviere 1992) appeals to applied researchers for a number of good reasons:

- It presents respondents with a simple and theoretically appealing task
- By constraining responses, it prevents scale use bias, which makes it ideal for supporting segmentation and allowing cross-cultural comparisons (Cohen 2003)
- It produces sensitive and discriminating measures (Chrzan and Golovashkina 2006)
- Available commercial software makes design and analysis of maxdiff experiments easy (Sawtooth Software 2005).

If one uses the Sawtooth Software maxdiff software to make a design rather than a balanced incomplete block design, one must decide how many maxdiff items (attributes) to include in each choice question. Bryan Orme (2005) addressed this question with an analysis of synthetic data. He found that a substantial increase in predictive accuracy occurs in maxdiff experiments with 5 items/set rather than 3 items/set. A smaller improvement occurs when moving from 5 to 7 items/set, however.

We seek to augment Orme's findings with analysis of data from real respondents. For our empirical analyses we draw data from three commercial studies, in each of which we systematically manipulated the number of items per question across cells of otherwise similar respondents.

Planned Comparisons

Dropout Rates
We use χ^2 tests to assess whether respondents quit surveys at different rates when maxdiff questions have more or fewer items per set. Because the dropout rates we measured were for the whole survey, and not just for the maxdiff section, we expect this to be a weak test.

Task Length
Timers at the beginning and end of the maxdiff portion of each questionnaire allow us to measure the duration of the maxdiff questions. Using a median test (because some respondents pause the survey during the maxdiff section) we test for the effect of the number of items per set on task length. With a regression analysis we quantify the contribution of the number of maxdiff questions and the number of items per question on task length.

Positional Bias
Order or position effects may be more pronounced as the number of items per set increases or decreases, and we track these to check.

Parameter Equivalence
Using the Swait and Louviere (1993) procedure for separating the scale from substantive parameters, we test whether maxdiff question with different numbers of items per set result in substantively different model parameters, or in models with different amounts of response error (scale).

Predictive Validity
Taking Elrod's (2001) advice against relying on hit rates to validate our maxdiff models, we supplement hit rate analysis with more meaningful out of sample validation tests.

Empirical Studies

Study 1
As part of a 16 minute interview about vacationers' travel preferences, 884 respondents evaluated the appeal of 17 activities in maxdiff questions. Each respondent completed 17 maxdiff questions, as follows:

- 220 had maxdiff questions containing 4 activity items
- 223 had maxdiff questions containing 5 activity items
- 220 had maxdiff questions containing 6 activity items
- 221 had maxdiff questions containing 7 activity items

For a validation holdout task we had respondents rank six of the 17 objects known from prior research to span the spectrum from low to high preference.

Study 2
Like Study 1, this study of 19 activities formed part of a larger study of vacation preferences. A total of 1,236 respondents divided randomly and evenly into three treatments:

- 19 questions, 3 items/question
- 19 questions, 5 items/question
- 19 questions, 8 items/question

Again respondents ranked 6 of the activities that spanned the preference range.

Study 3
This smaller study tested just 12 activities, again as part of a larger piece of travel research. A total of 904 respondents completed 12 maxdiff questions as follows:

- 302 had 3 items in each maxdiff question

- 300 had 5 items/maxdiff question
- 302 had 7 items/question

In this case the client cancelled the holdout question, so we split the sample in half and use the maxdiff model from one half to predict best and worst choices in the other half, and vice versa, as our holdout strategy.

Results

Dropout Rate
In study 1, more items per question produced higher dropout rates, results strong enough for a marginally significant linear trend ($p<.11$). In Study 2, maxdiff questionnaires with 3 items per set had lower dropout rates than those with either 5 or 9 items per question. Study 3 however, had much lower dropout among respondents receiving 5 items per question than among those receiving 3 or 7.

	% Dropout		
# Items/set	Study 1	Study 2	Study 3
3	-	2.6	3.2
4	10.0	-	-
5	13.5	6.8	0.7
6	13.6	-	-
7	14.9	-	3.8
8	-	5.5	-

As expected the dropout rate differences only weakly point away from using larger numbers of items per question.

Task Length
In all three studies, highly significant ($p<.001$) χ^2 tests for median test time showed longer task length as the number of items per questions (and the number of questions) increased:

	Interview Length (seconds)		
# Items/set	Study 1	Study 2	Study 3
3	-	207	95
4	220	-	-
5	248	292	139
6	279	-	-
7	390	-	138
8	-	361.5	-

A handy regression equation summarizes the impact of number of questions and number of items per question on task length:

Length (sec) = 9.4(number of questions) + 17.5 number of items/ question)

Positional Bias

Very slight position effects seem to occur, but they do not differ based on the number of items per question. For example, in Study 3, a slightly greater number of bests seem to occur in the first position, and a slightly greater number of worsts seem to occur in the last two positions, but (a) the effect is minimal, and (b) it is similar regardless of the number of items per question:

Position	3 items		5 items		7 items	
	b	w	b	w	b	w
1	36%	33%	24%	16%	16%	13%
2	36	33	19	21	16	15
3	29	34	22	22	15	13
4	-	-	18	24	13	13
5	-	-	16	18	15	14
6	-	-	-	-	15	17
7	-	-	-	-	11	15

Parameter Equivalence

Earlier work assessing the number of question to include in partial profile choice questions (Patterson and Chrzan 2003) found equivalent substantive parameters across partial profile questions of different sizes, but more response error (a smaller scale factor) for question with more attributes per profile. In the case of maxdiff scaling, we expected to find this but we did not: In Study 1 each of the 6 pairs of models had significantly different substantive parameters. Because the Swait and Louviere (1993) procedure is sequential and allows a test of scale only if the test for substantive parameter differences is non-significant, this rendered us unable to test scale parameters. These scale effects were almost negligibly small even if we could not test them for significance. What at first appeared odd became the norm when the same outcome occurred in Studies 2 and 3.

A glance at the aggregate MNL parameters from Study 3, however, reveals that while the substantive parameter vectors may be significantly different, the practical differences among them are small, and would not affect decisions made on the basis of the research:

Item	3 items/set	5 items/set	7 items/set
1	.60	.73	.74
2	-.69	-1.01	-.92
3	.14	.29	.09
4	.87	.99	1.02
5	.04	-.12	-.16
6	-.30	-.30	-.21
7	.07	.08	.11
8	-.63	-.61	-.53
9	-.63	-1.03	-.99
10	.59	.63	.57
11	.53	.49	.54

Studies 1 and 2 also had similar, but statistically significantly different, parameter vectors.

Predictive Validity
Hit rates are a poor man's measure of predictive validity (Elrod 2001), but we report them for their familiarity. Neither prediction of first choice nor of full rank orders differs significantly in quality depending on the number of items per question in Study 1:

Items/set	First Choice %	Rank Order %
4	63.6%	49.9%
5	68.6	54.2
6	60.0	52.0
7	62.4	52.0

In Study 2, first choice hit rates improve with increasing numbers of items per question ($\chi^2 = 13.6$, $p < .01$) but this pattern disappears for the prediction of the full rank order:

Items/set	First Choice %	Rank Order %
3	64.1	47.8
5	72.1	49.8
8	75.5	49.4

Root mean square error (RMSE) is a measure of how well predictions of shares match actual shares: higher RMSE is worse than smaller. Both for holdout samples of a rank order (Studies 1 and 2) and of best and worst choices (Study 3) RMSE shows no significant differences by number of items per question, though there is a hint that 3 items per question may produce slightly worse predictions than questions with more items:

Items/set	Study 1	Study 2	Study 3
3	-	4.1	13.1
4	3.1	-	-
5	3.8	2.9	12.2
6	5.5	-	-
7	4.7	-	11.1
8	-	4.7	-

Discussion

No qualitative differences in results emerge from these analyses as a function of the number of items per question. Given Orme's (2005) findings about increasing accuracy with larger numbers of maxdiff questions, and our finding that task length increases with number of items per question, we recommend a larger number of questions with smaller numbers of items per question. Given the slight evidence of poorer hit rates and poorer out-of-sample for 3 items per question we recommend using 4 or 5 items per question in maxdiff experiments. For example, in the same amount of time, respondents could rate

20 maxdiff questions with 4 items per question or 15 with 7 items per question. With mixed evidence about the number of items per set (except that smaller questions are shorter questions) and Orme's evidence of greater accuracy with more questions, the design having respondents complete 20 questions with 4 items per question makes more sense than the design with 15 questions and 7 items per question.

References

Chrzan, Keith and Natalia Golovashkina (2006) "An Empirical Test of Six Stated Importance Measures," *International Journal of Market Research*, in press.

Cohen, Steve (2003) "Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation," *2003 Sawtooth Software Conference Proceedings*, Sawtooth Software, 61-74.

Elrod, Terry (2001) "Recommendations for Validation of Choice Models," *2001 Sawtooth Software Conference Proceedings*, Sawtooth Software, 225-43.

Finn, Adam and Jordan Louviere (1993) "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety," Journal of Environmental Economics and Management **29**, 228-37.

Orme, Bryan (2005) "Accuracy of HB Estimation in MaxDiff Experiments," Sawtooth Software Research Paper, http://www.sawtoothsoftware.com/download/techpap/maxdacc.pdf

Patterson, Mike and Keith Chrzan (2003) "Partial Profile Discrete Choice: What's the Optimal Number of Attributes," *2003 Sawtooth Software Conference Proceedings*, Sawtooth Software, 173-85.

Swait and Louviere (1993) "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models," *Journal of Marketing Research*, **30**, 305-14.