Sawtooth Software

RESEARCH PAPER SERIES

A Parameter Recovery Experiment for Two Methods of MaxDiff with Many Items

Keith Chrzan,
Sawtooth Software, Inc.

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

A Parameter Recovery Experiment for Two Methods of MaxDiff with Many Items

Keith Chrzan, Sawtooth Software
January, 2015

Background

Clients don't seem to be able to get enough of a good thing and this seems to apply more to MaxDiff than to some of the other methods we use: clients frequently ask for MaxDiff experiments that include more items than would allow us to expose each item to each respondent the recommended three or four times.

In a paper at a recent Sawtooth Software Conference, Wirth and Wolfrath (2012) introduced two methods for handling large numbers of items in MaxDiff studies. "Express" MaxDiff creates different subsets of the large number of items and then asks a given respondent MaxDiff questions that just include that subset of items, with each respondent seeing each of the reduced number of items the recommended three to four times and with different respondents seeing different subsets of attributes. Express MaxDiff relies upon the magic of HB analysis to fill in the blanks and supply utilities for the items missing from a given respondent's experiment (essentially imputing them based on the population means and covariances). With "Sparse" MaxDiff, on the other hand, each respondent sees all the items in the study, but fewer than the recommended number of times (i.e. perhaps just once). In an empirical study Wirth and Wolfrath compared the ability of Express and Sparse MaxDiff to predict a few holdout questions: they found that both methods predicted "best" choices about equally but that Sparse did a better job predicting "worst" choices.

Wirth and Wolfrath (W&W) also conducted a parameter recovery experiment for Express MaxDiff but not for Sparse MaxDiff. An experiment comparing the ability of Express and Sparse MaxDiff to recover known utilities would help us understand which of the two works better. The following sections detail two such experiments using data sets whose owners allowed sharing of the data.

Study Methodology

The first study had 84 items (and 729 respondents) and the second 90 items (and 200 respondents). Using the HB utilities from these studies as the "known" true utilities grounds the artificial data experiment as realistically as possible in that it retains the realistic statistical properties (means, covariances among attributes) that one would see in living human respondents. The MNL model that underpins statistical analysis of MaxDiff experiments has strong assumptions that we can use to apply a theoretically appropriate pattern of random response error (one that conforms to the Gumbel distribution). Moreover, we can add the right amount of response error by making sure our artificial respondents have about the same level of test-retest reliability we see in human respondents. The final step in generating the artificial responses is to design the Express and Sparse questions and then have the artificial

respondents make their choices. The first study featured W&W's original method for making the Express MaxDiff design while the second used SSI Web's constructed list capabilities to generate a random subset of attributes for each respondent's Express MaxDiff design. In the first study artificial respondents chose the highest and lowest (utility + positively-skewed Gumbel error) alternatives from each of the MaxDiff questions. The second study applied the errors even more carefully and used positively-skewed Gumbel error for the "best" choices and negatively-skewed Gumbel error for the "worst" choices. Doing all this involved using some fancy perl script and the data generator capability in SSI Web.

The final step, running MaxDiff HB analysis on both resulting data sets, allows us to take the "true" utilities from the actual respondent data and then see how the Sparse and MaxDiff would be able to recreate those true utilities after being perturbed by random error and run through the two experimental designs.

Results

At the aggregate level (looking at mean HB utilities) we can compute the correlation of the mean utilities calculated via each Express and Sparse MaxDiff with the mean true utilities (from human respondents) and we can test the difference in these correlations using a t-test for differences in dependent correlations (Cohen and Cohen 1983). In both studies the correlation between known and estimated aggregate utilities shows that the two methods perform very similarly: 0.993 for Sparse and 0.985 for Express in the first study and 0.996 for Sparse and 0.992 for Express in the second. While trivially small and of no practical value, these differences are statistically significant ($p<0.001$). If you only need mean utilities, either method works great.

Often, however, aggregate utilities will not allow us to meet a study's objectives and we need to rely upon the quality of respondent level utilities (for e.g. simulations, TURF, segmentation). To assess the quality of respondent level utilities we can calculate the correlations of Express and Sparse HB utilities with the known utilities for each respondent; with paired correlations we can run a dependent t-test for means as we could with any other paired variables. Looking at the correlations at the respondent level the average correlations across respondents of Sparse/Express with actual utilities were 0.752/0.684 for the first study and 0.855/0.802 for the second. In this case the differences between Sparse and Express in both studies are large enough to be meaningful as well as highly significant ($p<.001$). For studies requiring respondent level utilities, Sparse appears to be a better way to go than Express (though both approaches involve sacrificing individual-level precision compared to typical MaxDiff studies wherein each respondent sees each item multiple times).

Discussion

In the Sawtooth Software User Group discussion on LinkedIn that followed the overview of this research and in private replies to the post, several folks made useful observations and suggestions, among them:

- First, the RLH statistic that some analysts use as a basis for determining respondent quality (and potentially for excluding respondents from reporting) will be more reliable with Express than with Sparse.

- Joel Anderson had a clever suggestion: he thought that, better than using artificial respondents one could just take the data from actual respondents and discard a random 2/3 of their choices, effectively giving them something between a Sparse and an Express design. Joel found that the method reproduced the utilities from the full data set very well. A redo of this experiment with very controlled selection of the choice sets to discard might well allow a rigorous test of Express and Sparse MaxDiff using data from human respondents, something well worth sharing at one of our conferences (hint, hint).

- Tom Eagle pointed out that measurement theory might suggest that Express would outperform Sparse: For a given item Express MaxDiff gets really good information on 30 out of (say) 90 items and imputes the value of the other 60 items while Sparse MaxDiff may get poor measurements of all 90 items.

- Several people noted that artificial respondents really are not real humans. Of course this is true: no matter how lovingly and realistically we craft artificial respondents, and even if we add theory-driven amounts and distributions of random error to their responses, we simply cannot guarantee that what we find with artificial respondents will generalize to human respondents (For example, real humans may benefit from seeing items multiple times, because when they have seen an item before they may spend less cognitive effort in subsequent viewings (as opposed to seeing each item just once fresh and new): hence the value of experiments along the lines that Joel Anderson recommends.

If you like this discussion, you may also want to check out the paper "Bandit Adaptive MaxDiff Designs for Huge Numbers of items" at the 2015 Sawtooth Software Conference in March 2015 in Orlando, Florida (Fairchild *et al.* 2015). That paper focuses on finding the top items in a long list (one of their examples includes 300 items!) rather than providing quality utilities for all the items in a study at the individual respondent level. Interestingly enough, their simulation results also suggest a slight edge in performance for Sparse over Express MaxDiff.

References

Cohen, J and P. Cohen (1983) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 2nd ed.* Hillsdale: Lawrence Erlbaum.

Fairchild, Kenneth, Bryan Orme, and Eric Schwartz (2015), "Bandit Adaptive MaxDiff Designs for Huge Number of Items," *2015 Sawtooth Software Conference Proceedings*, (forthcoming).

Wirth, R. and A. Wolfrath (2012) "Using MaxDiff for Evaluating Very Large Sets of Items," *2012 Sawtooth Software Conference Proceedings*, pp. 59-78.