Sawtooth Software

RESEARCH PAPER SERIES

Balancing Version Incidence in Administration of CBC and MaxDiff Surveys: Important or Not?

Keith Chrzan
Sawtooth Software, Inc.

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

Balancing Version Incidence in Administration of CBC and MaxDiff Surveys: Important or Not?

Keith Chrzan
Sawtooth Software, Inc.

October 2014

How important is it that equal numbers of respondents complete each of the several versions (or blocks) of a CBC or MaxDiff survey? If version balance did matter, then it could complicate CBC and MaxDiff studies in practice, particularly when a survey has quota samples for respondent segments or when we are fielding an international survey: we run into some administrative complexity if we need to have exactly 10 respondents for each of our 30 versions in each of our four user segments in each of eight countries.

Parameter Recovery Experiment for CBC

Design of the Experiment

One way to approach this question is to see how well we can recover the known utilities of a sample of respondents who answer a survey under a condition of balance (equal numbers of respondents per version of the survey) and then again under a condition of imbalance (unequal numbers of respondents per version of the survey). Unfortunately, we are never in position to know the true utilities of any set of human respondents because all we can collect from them are utilities refracted through a CBC experimental design and HB estimation. We can, however, run this experiment with realistic synthetic respondents.

To make our synthetic respondents as lifelike as possible, we use utilities from 300 respondents in a recent field experiment. This preserves the patterns of heterogeneity among respondents and covariances among the utilities of the various attribute levels. To this realistic base we add just such response error as produces a 65% test-retest hit rate (for the three-alternative choice tasks used in this study) – within the range we see in well-designed commercial studies. We simulate choices by summing the "true" utilities for each alternative in each choice set, adding independent and identically distributed Gumbel errors to each alternative and then identifying as chosen the alternative in each choice set with the highest sum of utility plus error.

We created a $5^2 \times 4^2 \times 3^2$ design with 30 versions of 10 tasks, each with three choice alternatives and forced choices (i.e. no "none" alternative). In our balanced condition 10 respondents complete each of the 30 versions of the survey. In the unbalanced condition, different versions receive different numbers of respondents, as follows:

Version	# of Respondents
1-6	1
7-12	4
13-18	10
19-24	15
25-30	20

Analysis Plan

After CBC/HB analysis of the 300 respondents under the two conditions we have three sets of utilities:

- The respondents' true utilities – the ones we used to generate the choices in both balanced and imbalanced data sets
- CBC/HB estimated utilities from respondents under the balanced condition
- CBC/HB estimated utilities from the same respondents under the imbalanced condition

We can measure parameter recovery by looking at the correlation of true with estimated utilities under both conditions and for each individual respondent. Moreover, we can subject the resulting correlations to statistical testing to see if one set is higher than another. With our sample size of 300 we have enough power to detect a non-trivial difference between conditions.

So what happens?

Results

Averaged across the 300 respondents, the correlation of known and estimated utilities is 0.827 for the balanced condition and 0.834 for the unbalanced condition. This difference is not significant at conventional levels ($Z=0.81$, $p=0.21$). In terms of parameter recovery, having a balanced or imbalanced number of respondents per version makes no significant difference.

Digging into the results a bit deeper, what if we compare the most extreme sets of versions – the under-represented versions (those with one or 4 respondents each) and the over-represented versions (those with 15 or 20 respondents each)? Again, however, we see no statistically significant difference in parameter recovery ($Z=1.25$, $p=0.106$).

Parameter Recovery Experiment for MaxDiff

Design of the Experiment

We can conduct a similar test for MaxDiff, again using realistic synthetic respondents (based on utilities from 396 human respondents). The MaxDiff experiment includes 20 items, 27 versions and 15 sets of four items each per version. The MaxDiff experiment may be a little less realistic than the CBC experiment because we assume that the same utility function governs the choice of "worst" items as of "best" items, though we know that the two sets of choices often generate slightly different utility vectors.

In the balanced condition 14-15 respondents complete each of the 27 versions of the survey. In the unbalanced condition, different versions receive different numbers of respondents, as follows:

Version	# of Respondents
1-3	1
4-6	5
7-9	9
10-12	12
13-15	15
16-18	18
19-21	21
22-24	24
25-27	27

Results

The results for MaxDiff show even less impressive differences in utility recovery under different conditions of version balance. Mean correlations with the "true" utilities are 0.918 for the utilities estimated under version balance and 0.919 for utilities estimated under version imbalance ($Z = 0.29$, $p = 0.39$).

Conclusion

The two experiments, each involving a more extreme case of version imbalance than is likely to occur in practical applications, provide no evidence that version imbalance will produce utilities significantly better (or worse) than would result from having versions with equal numbers of respondents. Parameter recovery does not differ significantly under conditions of balance or of imbalance, likely because we strive for level balance and orthogonality within each version. The designs in SSI Web are not like the blocked, fixed designs one would pull from an experimental design catalog, where the blocks are like puzzle pieces and missing any of them ruins the whole: because each individual version of the experiment stands alone well in terms of being nearly level balanced and orthogonal, even if some versions have no respondents at all, the quality of the overall will not be greatly affected.

Of course these studies use synthetic respondents. While we went to great pains to make respondents as human as possible (preserving respondent heterogeneity and utility covariances while allowing for a realistic level of respondent imperfection), it remains possible that the results from a simulation experiment may be somehow different from what would happen with real human respondents. Possibly, some aspects of human choosers (e.g. sensitivity to order effects) not programmed into our synthetic respondents could produce different results than those reported above, though using a large number of versions and having even semi-good balance across them should counteract such order effects. As noted, this caveat may be more important for the MaxDiff test than for the CBC test because we know that the utilities for Worst choices are not just the utilities for Best choices "upside down and backwards in a black mirror." However, a study of version balance featuring human respondents would seem to introduce additional potential sources of error (e.g. are respondents in the two cells really equivalent?). But perhaps a clever reader will figure out how to address these challenges and bring an empirical treatment of this topic to one of our conferences!