Predictive Validity of Conjoint Analysis Results based on Best-Worst Scaling compared with Results based on Ranks Data

Hume Winzar, Griffith University
James Agarwal, University of Calgary
Barbara Khalifa, Liane Ringham, Inside Story Research & Knowledge Management

Abstract

The time-consuming nature of Best-Worst deriving scales for a conjoint analysis study leads the authors to question if ranked data, which could be translated into equivalent B-W scores, might be nearly as useful. Respondents answered a B-W scaling task regarding a familiar, highly involving product category, and then two-weeks later did a sorting task on the same fractional factorial array. Conjoint analysis coefficients were calculated and values were forecast for four holdout profiles. Kendall's tau measured the extent that the holdout ranks were reproduced in conjoint predictions. Results suggest that overall the rank data were comparable to the B-W data, but with greater variance.

Background

Different data gathering approaches have been used in Conjoint Analysis research. These are discussed in Louviere (1994). One approach that is gaining in popularity is Best-Worst scaling. Finn & Louviere (1992) describe Best-Worst scaling as a multiple-choice extension of Thurston's method of paired comparisons. Best-Worst differs from ranking scaling in that respondents evaluate objects in sets of incomplete subsets of all objects, whereas in ranking all objects are evaluated together.

Key advantages of B-W measurement over rating scales are that it is scale-free, so it is ideal for application across different cultural groups which use scales quite differently (Cohen and Neira 2003), no prior assumptions need to be made about the scaling of evaluation and choice and, unlike ranking tasks, Best-Worst allows for ties in evaluations and for skewed preference functions. Other claims are that respondents spend more time on a Best-Worst task and their evaluations are more well-considered and accurate. Some of these claims have considerable empirical support (e.g. Marley and Louviere 2005 on scale-free properties). Others are as much supposition or extrapolations of what we understand from existing preference elicitation procedures. Some of the frequently cited references on the nature of Best-Worst scaling appear in unpublished working papers that seem to be no longer available or in circulation (e.g. Louviere 1991; Louviere, Swait & Anderson 1995).

Example Best-Worst Measurement

Best-Worst scales are derived using the following procedure. Say a fractional factorial array for a conjoint analysis study gives us eight profiles for evaluation, as in Table 1. A balanced incomplete block (BIB) design of 14 blocks of 4 is feasible, or a partially balanced incomplete block (PBIB) design of just six evaluation sets of four profiles can be extracted as shown in Table 2. Thus each profile is evaluated three times, and against most other profiles three

times. Respondents are given a block of profiles, such as in Table 3, and asked to select the best option and the worst option. This procedure is repeated 5 more times to gather a complete set of Best-Worst evaluations. Each profile then is simply scaled as the number of times it was selected as best minus the number of times selected worst. In the PBIB used here a profile would then have a score ranging from 3 to -3 (3 for best three times and worst none, to -3 for best none and worst all three times.)

Table 1: Sample Fractional Factorial Array for Conjoint Analysis study

#	Profile
1	Club Embarco: This 4-star hotel located right on the beach, and 20-minutes from Puerto Vallarta's biggest nightclubs. All inclusive CN$1472.
2	Hacienda Blanca: This 3-star hotel located right on the beach, and 10-minutes from Puerto Vallarta's best nightclubs. All inclusive CN$1278.
3	Club del-Rio: This 3-star hotel located right on the beach, and 20-minutes from Puerto Vallarta's biggest nightclubs. All inclusive CN$1715.
4	Casa Puerto: This 4-star hotel located just 20-minutes from Puerto Vallarta's best nightclubs. Despite its off-beach locale, the hotel has a private beach-club on the ocean only a ten-minute walk away. All inclusive CN$2019.
5	Villa Vallarta: This 3-star hotel located just 20-minutes from Puerto Vallarta's biggest nightclubs. Despite its off-beach locale, the hotel has a private beach-club on the ocean only a ten-minute walk away. All inclusive CN$1472.
6	Casa del-Mara: This 3-star hotel located right on the beach, and 10-minutes from Puerto Vallarta's biggest nightclubs. All inclusive CN$2019.
7	Club Madrid: This 4-star hotel located just 10-minutes from Puerto Vallarta's best nightclubs. Despite its off-beach locale, the hotel has a private beach-club on the ocean only a ten-minute walk away. All inclusive CN$1278.
8	Hacienda Pisco: This 4-star hotel located just 10-minutes from Puerto Vallarta's best nightclubs. Despite its off-beach locale, the hotel has a private beach-club on the ocean only a ten-minute walk away. All inclusive CN$1715.

The Problem

In the simple example presented here it is fairly easy to derive the Best-Worst measures. A larger number of profiles with a properly balanced incomplete block design could present a very large number of evaluation sets, requiring multiple respondents (losing the advantages of individual-level data) or risking impatience from respondents.

The authors were confronted with this problem in an applied research setting. A possible compromise would be to ask respondents to simply rank the options instead of making a succession of B-W judgements. Ranks could be easily transformed to the same type of scale as the B-W scales. For example considering the eight profiles in our example, we can interpret ranks as the relative position of each profile against each other profile giving n-1=7 judgements per profile. The single most attractive option would win all seven times and lose never. The second most attractive option would win six of the seven times evaluated, and lose once (to the most attractive option.) Thus the ranking of eight profiles would give a B-W transformation ranging from 7 to -7 in increments of 2.

Best-Worst scaling was the better approach with respect to capturing accurately the utility functions amongst respondents. B-W permits skewed utility and indifference in attractiveness

of some product attributes, or product profiles. On the other hand, it is very time consuming and potentially fatiguing, although evidence shows that fatigue alone does not affect the reliability of a choice experiment (Swait and Adamowicz, 2001). We wondered if it would make much difference if we simply asked respondents to rank the options.

Table 2: Partially Balanced Incomplete Block Design for 8 Profiles

Block #	Profiles			
1	1	5	2	6
2	1	5	3	7
3	1	5	4	8
4	2	6	3	7
5	2	6	4	8
6	3	7	4	8

Table 3: Example: Block #1 for Best-Worst Evaluation (profiles 1, 5, 2 & 6)

Best		Worst
	Club Embarco: This 4-star hotel located right on the beach, and 20-minutes from Puerto Vallarta's biggest nightclubs. All inclusive CN$1472.	
	Villa Vallarta: This 3-star hotel located just 20-minutes from Puerto Vallarta's biggest nightclubs. Despite its off-beach locale, the hotel has a private beach-club on the ocean only a ten-minute walk away. All inclusive CN$1472.	
	Hacienda Blanca: This 3-star hotel located right on the beach, and 10-minutes from Puerto Vallarta's best nightclubs. All inclusive CN$1278.	
	Casa del-Mara: This 3-star hotel located right on the beach, and 10-minutes from Puerto Vallarta's biggest nightclubs. All inclusive CN$2019.	

The Experiment

A convenience sample of final-year undergraduate students at a Canadian university was contacted about four weeks before "Reading Week" in March 2007. Reading week, like the US "Spring Break" is supposed to be a chance to catch up with studies but is traditionally an opportunity to visit exotic places for a week-long party. Canadian students frequently take package tours to the Mexican cities of Cancun or Puerto Vallarta. The product category then was one with which all students were familiar and many had some experience. Several students were about to return for a fourth time.

Respondents completed a Best-Worst scaling task using the partially balanced incomplete block design summarised in Table 2, in the format shown in Table 3, after being given the following scenario:

- "You and some friends have decided to go to Porto Vallarta for Reading Week. Some “last-minute” packages have just become available. All packages are the same (all-

inclusive meals, open bar, air-fares, transfers and taxes) except for the quality of the hotel, distance from the beach and from nightlife and, of course, prices… "

Immediately after the Best-Worst task, respondents were asked to examine four holdout profiles that were not in the original mix and asked to rank them most attractive to least attractive. Two weeks later, the same student group was asked to examine the same profiles in a shuffled set of cards. Respondents then sorted the cards into rank order from most attractive to least attractive. The ranked data were then translated into equivalent B-W scale values. Respondents were debriefed on the purpose of the research and appropriately thanked for their contribution.

To complete the Conjoint Analysis procedure, scores for the B-W measures and the B-W transformed ranks were decomposed using OLS regression to extract utility coefficients for each of the attribute levels - hotel quality, distance from beach, distance from night clubs and price. Using these utility coefficients forecast values for the four holdout profiles were calculated.

Results

A total of 37 respondents completed both B-W and sorting tasks over the two-week period. Interestingly only one of those respondents had a B-W score that corresponded with a clear rank 1-thru-8, reinforcing the view that few evaluation tasks are completely unidimensional.

Of interest to us here, however, is not whether ranks are the same as B-W (clearly they're not) but whether rank data can give us input suitable for a conjoint study that are as useful as those from B-W data. For each person in the study the ranks for the four holdout profiles were compared with the forecast scores derived from the conjoint analysis predictions from both the B-W task and the rank task. Results are summarised in Figure 1.

We can see that the predictions based on the rank data compare favourably with those of the B-W data. Average correlation is not much different from that of the B-W result, and the rank results feature more "perfect" correlations. The key difference is the range of values: the variance for the rank data is much greater than for the B-W data, there are two values that are negative, indicating that forecast values are in the reverse order than the true values, and there are several other very low correlations.

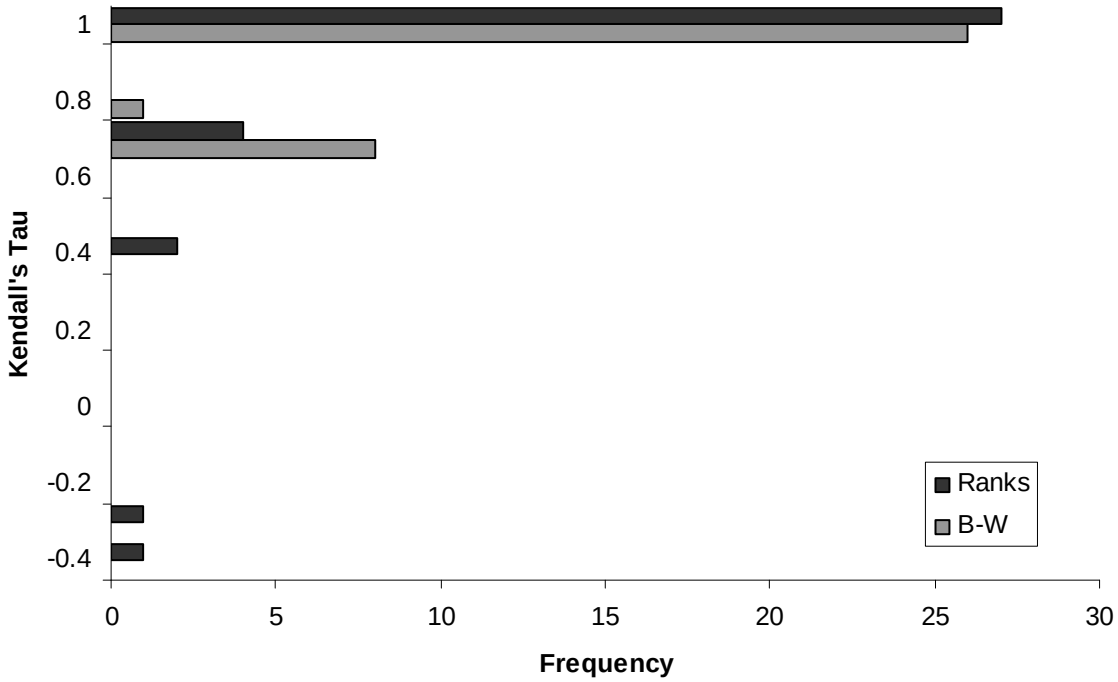
Conclusions

The purpose of this study was to test the feasibility of using ranking data as input for a conjoint analysis study instead of the more reliable Best-Worst scaling method. Results in this small study suggest that it would be a quite sound method.

Some caveats need to be placed on studies using such a shortcut, however. Some method for capturing those people with inconsistent evaluations, such as those with negative rank correlations, may be appropriate. This study used a sample from a population that was familiar and highly involved with the product category, the profiles were fairly simple and neither of the evaluation tasks was difficult or time-consuming. We cannot draw conclusions from this study about complex, low-involving or unfamiliar products. In this study B-W was

conducted first, followed by sort. While a two-week break probably removed priming effects, a two-group experiment would be more valid. We used a Partial BIB in this study because it was considerably shorter than the next-best 14-block BIB. It is possible that the BIB will give different results. This experiment may have favoured one procedure over the other by gathering the holdout data shortly after the B-W task. Finally, rank correlation may not be the best measure for applied research. Sales and share prediction more often are the goals, so these may be preferred.

Figure 1: Kendall's Tau: Best-Worst conjoint predictions and Transformed Rank predictions against four holdout profiles.



	B-W	Rank
Mean	0.88	0.83
Min	0.55	-0.67
Max	1.00	1.00
StDev	0.15	0.38

-0-

References

Cohen, Steve and Leopoldo Neira .,2003. Measuring Preference for Product Benefits Across Countries: Overcoming Scale Usage Bias with Maximum Difference Scaling (ESOMAR 2003)

Finn, Adam and Jordan J. Louviere., 1992. “Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety,” Journal of Public Policy and Marketing, Vol. 11, No. 2 pp.12-25.

Louviere, Jordan., 1991. Best-Worst Scaling: A Model for the Largest Difference Judgements, Working Paper, University of Alberta, Canada.

Louviere, Jordan., 1994. "Conjoint Analysis." In R. Bagozzi (ed.), Handbook of Marketing Research. Oxford: Blackwell.

Louviere, Jordan, Joffre Swait, and Donald Anderson., 1995. Best/Worst Conjoint: A New Preference Elicitation Method to Simultaneously Identify Overall Attribute Importance and Attribute Partworths, Working Paper, University of Sydney, Australia and Working paper, University of Florida, Gainesville, FL.

Marley, A.A.J. and Louviere, J. J., 2005. Some probabilistic models of best, worst, and best–worst choices, Journal of Mathematical Psychology, Vol. 49, No. 6, pp.464-480

Swait, J. and Adamowicz, W., 2001. The influence task complexity on consumer choice: a latent class model of decision strategy switching, Journal of Consumer Research, Vol. 28, pp135-148.