Journal of Business Studies, Vol. XXXVII, No. 1, April 2016

Review of Best-Worst Scaling Method: A New Method over other Scales in Marketing Research

Shehely Parvin*

Abstract*: The traditional approach is to quantify key factors in marketing research through the use of universal rating scales. However, the use of rating scales does not always lead to good findings because of various response style biases such as social desirability bias, acquiescence bias, and extreme response bias. The Best-Worst scaling approach may be a method to overcome these problems by asking respondents to make trade-offs among the variables being assessed. This paper provides a comprehensive review of the theoretical and statistical underpinnings of BWS and a step-by-step managerially usable method which requires not much of training, statistical skills, or special software. The paper also reviews on the limitation of this technique and applications of the approach.*

Keywords*: Best-worst scaling method; rating scale; choice set.*

Introduction

When conducting research using surveys, a researcher could develop the data collecting procedure in different ways such as rating-based models, ranking tasks, constant sum tasks, paired comparison methods (Bednarz, 2006). Among them rating scale is mostly common used method in marketing research. One classic form of a rating scale is the Likert-type scale where the subjects are asked to tick their rating for each characteristic. Researchers frequently use adjectival descriptors to label the scale categories (e.g. “important”, “not important”, “good”, “fair” or “strongly agree”, “neutral”, “strongly disagree”) and follow unlike rating scale such as 3-point, 4 point, 5 point or 7 point or 9 point in developing scale for surveying. Henceforth, meaning associates to these different adjectival descriptors on different interval scales may be different from one respondent to another’s. As a consequence, there is less possibility to deduce responses. Accordingly, the conclusions based on rating scales could be biased to researchers. In this regard, the Best-Worst scaling method permits the cognitive process by which respondents identify the two items with the most and the least of a characteristic from designed subsets of three or more items. And it has several advantages that overcome the limitations of other methods of measurements.

Background

Fundamentally, Best-Worst Scaling (BWS) method, also known as maximum difference scaling (Cohen 2003), was first proposed by Louviere and Woodworth (1990), and the formal statistical and measurement properties were proven by Marley and Louviere (2005). BWS is rooted in the well-established random utility theory in psychology (Thurstone 1927) and economics (McFadden

*Associate Professor, Department of Marketing, Faculty of Business studies, University of Dhaka, Dhaka 1000, Bangladesh.

1986). The BWS method is a comparatively new method of measurement that has a number of advantages in any research study (Louviere et al. 2013). Basically, the BWS method is a choice modelling experimental procedure that requires a list of attributes that need to be expressed as having a particular magnitude along some kind of continuum (such as 'importance') (Finn & Louviere, 1992). The BWS method effectively permits respondents to evaluate all pairwise combinations of alternatives presented in a particular subset leading to the assumption that their 'best' and 'worst' choices represent the maximum difference in utility between all attributes. Therefore,the BWS method has been found to achieve comparatively the most accurate and reliable data which has provided researchers with the highest level of discrimination between variables, thus having a higher tendency to predict what they are intended to predict (Cohen, 2003).

The BWS method operates under the assumption of Random Utility Theory (RUT) in psychology (Thurstone, 1927) where it is expected that respondents will give effort to find the pair of items that denotes the maximum difference in utility within a particular subset. The literature has pointed out that under the RUT, utility is apparent through the operation of the experiment, with the greatest utility in a subset being able to be broken down into observable and error components (Auger, Devinney & Louviere, 2007). The choice made can be identified as reflecting the respondent's decision strategy that represents the systematic component of utility, plus a component that reflects all unobservable/ unexplainable impacts on the choices made (Auger et al., 2007). Accordingly, by choosing the best/most and worst/least attributes repeatedly across a number of similar subsets, the relative probabilities can be predicted across a particular set of items.

As a consequence, the BWS approach has a number of methodological benefits that are useful for any research. It helps respondents to make trade-offs among the items being evaluated that lead to greater discrimination between the items being assessed, and a greater level of predictive validity (Cohen, 2003). Moreover, the BWS method creates a uni-dimensional interval-level estimate of the attribute levels, based on nominal-level choice data (Massey, Wang, Waller & Lanasier, 2013). Because BWS requires respondents to choose themaximally different pair (Cohen 2003; Cohen and Markowitz 2002), they cannot use the middle points, end points, or one end of a scale, thus minimising chances of the various types of response bias mentioned earlier. Consequently, the BWS technique is the best solution to solve the problems of 'end-piling' related to ratings measurements. Moreover, the BWS method asks the same thing multiple times in comparison to the rating method, thus increasing the reliability of the test: it has also been proven to take considerably less time for respondents to complete the questionnaire than the rating task (Lee, Soutar, and Louviere 2008). Furthermore, acquiescence and extremity response biases are also reduced in comparison to traditional rating scales as the construction of the best–worst 'subset' does not allow respondents the opportunity to distort their true choice (Lee, Soutar, and Louviere, 2007).

In addition, the BWS method is an easier task for respondents than ranking. Respondents frequently find it problematic to rank more than seven items in a ranking task: as a result, the test-retest reliability of long lists of ranked items tends to be low (Chapman & Staelin, 1982; Lee et al.,

2007). Furthermore, prior studies have indicated that cross-cultural equivalence using other methods, mostly rating scales, is very difficult to achieve (Lee et al., 2007). In contrast, the BWS method only permits the selection of the best/most important item and worst/least important item; therefore, it can reduce biases due to differences in cultural response styles. In this regards, Auger et al.'s (2007) study of consumer ethical beliefs across cultures delivers empirical evidence that supports the ability of the BWS method to decrease such response biases. A major strength of selection under BWS conditions is that not only are respondents faced with a cognitively simpler task than ranking, but their task also closely reflects how people actually make choices, i.e., respondents only consider the most distinct or maximally different pair rather than pickinga response along a continuum such as a Likert-type rating scale (Cohen and Neira 2003; Flynn et al. 2007).

The BWS permits researchers to use large samples, well above those which could genuinely be achieved using either focused group discussion, or expert panels. Consequently sampling error could be significantly reduced, further increasing the validity of the results. It is also worth mentioning that researchers and practitioners can also use this technique on samples that are smaller than would be required using traditional survey-based approaches using Likert scales. Finally, BWS has also been presented to take significantly less time for respondents to complete than the rating task (Lee et al., 2008). For these reasons, Almquist and Lee (2009) recommend that companies planning new product development in different market segments could take on BWS as a more reliable technique to find out what customers really want.

In summary, the BWS method offers the chance of a new theoretically valid method of data collection and it has also been confirmed to be easy for respondents to understand in comparison with other methods such as rating scales and ranking workouts (Chrzan & Golovashkina, 2006). The BWS questions have been proven to be simple and easy to complete and do not require too much thought or knowledge to undertake them (Flynn et al., 2007). In addition, the BWS method has been proven to have relatively low financial costs in its administration that, in turn, can boost managerial practicalities for the use of this scaling method in any situation (Finn & Louviere, 1992). For that reason, the BWS method has been applied in a wide range of contexts and to investigate a wide variety of problems. The BWS method was first introduced by Finn and Louviere (1992) to assess the relative importance of food safety against other areas of public concern. Marley and Louviere (2005) later offered formal mathematical proof relating to its measurement properties. The BWS method has since been applied in various contexts including studies in marketing and consumer behaviour (e.g. Auger et al., 2007; Louviere & Islam, 2008; Louviere et al., 2013; Massey et al., 2013); personality research (Lee et al., 2008); health economics (Lancsar et al., 2007); and education (Burke, Schuck, Aubusson, Buchanan, Louviere, and Prescott, 2013).

Theoretical Supports Indicating the Benefits of BWS over other Scales

In the following section the paper reveal the advantages of the BWS approach over other scales based on extant literature. For example, Lee et al., (2007), clearly compared this measurement approaches in their study of personal values known as Kahle's (1983) List of Values. A key outcome of their research was that rating scales lead to greater skewness in the data than BWS,

resulting in a positive bias in which respondents rate all of these personal values asimportant. Basically, personal values are inherently positive constructs, respondents often show little differentiation amongst them when measured using rating scales. McCarty and Shrum (2000) referred to this bias as 'end-piling' where respondents systematically respond positively to each personal value being measured, leading to an over-use of the higher response categories in the rating scale. A major problem with end piling is that it reduces the discrimination between the personal values being measured. This lack of discrimination can in turn affect the statistical properties of the values, and one's ability to detect their relationships with other relevant variables. In this regards, two prior studies have shown that BWS increases differentiation and reduces end-piling in the measurement of personal values (Lee et al., 2007; McCarty and Shrum 2000).

Similar type of results was found when BWS was applied to another well-known personal values instrument – Schwartz's Values Survey (Schwartz 1992; Schwartz and Bilsky 1987). Lee, Soutar, and Louviere (2008) shown that BWS yielded a far better fit to the theoretical quasi-circular structure of values proposed by Schwartz (1992) than ratingscales. For instance, in Schwartz's (1992) conceptualisation of 'universal values', those values which are most adjacent are most similar, e.g., power and security are likely to be compatible within the circular map of those values. In contrast, those values which are most different e.g., power and benevolence, are likely to be located on opposite sides of the quasi-circular structure (see Schwartz 1992, 45). This implies a set of positive and negative inter-correlations amongst these values. Using data collected from both the rating-based Schwartz Value Scale (SVS), and the Schwartz Values Best Worst Survey (SVBWS), the pattern of inter-correlations between the values using SVBWS showed greater discrimination than the raw SVS data. This suggests that the SBBWS method is better to the raw SVS scores, as it produced a correlation matrix which better reflected Schwartz's (1992) theoretical structure than did the raw SVS scores. Furthermore, a number of the values were out-of-place on the circular map when using rating-based SVS data. The values security and achievement should have been located on either sides of or at least adjacent to the value power (as was the case with the SVBWS map), but instead were located on the opposite side of Schwartz (1992) circle, closer to the value benevolence.

Besides this, earlier studies found it difficult to achieve cross-cultural equivalence using rating scales. This is due to the difficulty in finding lexically equivalent verbal descriptions for a rating scale, metrically equivalent distances between numbers, and separating numbers from their meanings (Craig and Douglas 2000; Mullen 1995). Since BWS only asks which item is the most important and which item is the least important, it is much more likely to reduce biases due to differences in cultural response styles. Auger, Devinney, and Louviere's (2007) study of consumer ethical beliefs across cultures deliver empirical evidence supporting the capability of BWS to reduce such response biases.

Correspondingly, a technical paper on BWS (i.e., Cohen 2003) compared the use of rating scales, paired comparisons, and BWS, and found BWS to be most discriminating, and monadic ratings least discriminating among 20 objects that were evaluated. Therefore, Cohen (2003) concludes that BWS is certainly a superior method of collecting preferences than a ratings task'. In recent times, Chrzan and Golovashkina (2006) compared BWS with five other importance measures (importance ratings, constant sum, Q-sort, unbounded rating scales, and magnitude estimation)

and found BWS beat the other measures with the greatest discriminating and predictive power. In fact, many empiricalinvestigations of traditional rating scale approaches versus BWS have revealed the supremacy of BWS across a wide range of contexts. Consequently, in the following section this paper reviews the existing literature to give a detailed methodology for conducting a BWS study.

Detailed Best–Worst Scaling Methodology

Designing the choice sets

Very beginning this approach needs to pre-specify the attributes/items of interest from the literature and attempts to measure those attributes. After that, the first step is to generate the 'choice sets', i.e., the specific combinations of three or more items from which respondents choose the most/least option. It can be determined by appropriate balanced in complete block design (BIBD) designs (Green 1974; Raghavarao and Padgett 2005). A BIBD design permits to greatly reduce the number of choice sets to be evaluated while maintaining balanced appearance and co-appearance of items across the sets. For instance, the total number of possible combinations of three items per choice set is given by the following Equation 1:

Total Possible Number of Combinations of Three-Item Choice Sets;

$$\frac{k(k-1)(k-2)}{6}$$

(Adapted: Messey et al., 2013)

Where: k = the number of itemsto be evaluated

In other situations yet it is not possible to use 3 items per choice set, and 4 are required to maintain the desired properties of BIBDs. The equation to estimate the total number of possible combinations using 4 items per choice set is given by Equation 2 (Green 1974; Raghavarao and Padgett 2005):

Total Possible Number of Combinations of Four-item Choice Sets

$$\frac{k(k-1)\times(k-2)\times(k-3)}{24}$$

(Adapted: Messey et al., 2013)

More specifically, for example, if nine personal values from Kahle (1983) have been listed for the BWS study then replacing in the known values into Equation 1 for the three-item choice sets leaves us with 9 (9 - 1) X (9 - 2)/6 = 84 possible combinations of those nine items, and for the 4-item choice sets in Equation 2, there are 9(8) X (7) X (6)/24 = 126 possible combinations. Clearly, asking respondents to evaluate 84 or 126 choice sets would cause fatigue, and jeopardise the validity of the study. On the other hand, auspiciously, when using BIBDs respondents do not need to evaluate all possible combinations (Cohen 2003; Cohen and Neira 2003). Instead they only need to be exposed to a limited number of choice sets provided by the BIBD, although this does not mean that the researchers relinquishes a significant amount of statistical information. In this

regards, to assist researchers and practitioners in using BWS, how to design a BWS study, a useful selection of BIBDs with varying numbers of items to be evaluated, and for 3 or more items per choice set are provided in Appendix A and B.

Therefore, with an appropriate experimental design, such as a balanced incomplete block design (BIBD) where items within the experiment are balanced, orthogonal and adequately randomised under the assumption of random utility theory (RUT) (Green, 1974), the error component of the utility of the maximum difference pair in the subset can be estimated. The major benefit of using a BIBD design is its capability of greatly decreasing the number of choice sets to be evaluated while maintaining the balanced appearance and co-appearance of items across the sets: the number of items that appear in each set ideally must be fixed at three or more (Green, 1974; Raghavarao & Padgett, 2005).

For instance, for a set of ***k*** items, a BIBD design will generate ***s*** choice sets; each choice set will have ***m*** items. To minimise the task difficulty, ***m***should always be less than the ***k*** items; each item appears ***r*** times and each pair of items appears **(λ)** times. In a BIBD design, no object appears more than once in a block; every pair of objects appears in the same number of blocks; each block is of equal size; and every object appears equally. A BIBD experiment must satisfy an integer's lambda value and ***r(m - 1)/(k - 1)*** equation will calculate the lambda value (Massey et al. 2013). If ***s*** is equal to ***k*,** the design is known as symmetrical BIBD (Raghavarao & Padgett, 2005). Whilst a symmetrical design is always favoured, it is not always arithmetically possible because of the three required restraints for this design. Therefore, there should be positional balance in an ideal BWS design that controls possible order effects with each respondent seeing each item in the first, second, third, etc. position across the sets (Lee et al., 2007). When the BIBD experiment is not symmetrical, it is required to randomise the order of items that are seen in each choice set to control for possible order effects (Massey et al., 2013).

Sample choice set for collecting the data for a BWS task

Respondents would ask to provide best/worst answers for the choice sets. A sample choice set is shown in Figure 2. As presented in the figure, respondents then ask to indicate the one they considered to be the most important values and the least important values in life.

Most Important (Tick ONE box)	**Of these, which are the most and least important?**	**Least Important (Tick ONE box)**
O	Being self-fulfilled in life	O
O	Having security in life	O
O	Having warm relationships with others	O

Figure 2: An example of Best-Worst Task
Adapted: Lee et al., 2007

Respondents then presents the items contained in choice set 2, and asks to indicate which amongst the new set of three personal values would be the most/least important. Though there is may be some inherent subjectivity in respondents' evaluations of the objects, the multiple measurements across the choice sets, and the multiple respondents would be used in the BWS task deliver a strong evidence-based set of evaluations. Therefore, these are more likely to be accurate than those generated using qualitative techniques or rating scale approaches. In this way, data collection would be completed when the respondents provide their best–worst responses for all choice sets.

Scoring Method of BWS Data

In regards of the Best–Worst Scaling (BWS) scoring procedure, the prior literature indicates that the overall score of this design can be calculated in many ways (Burke et al., 2013). Among them, one way is to calculate the best–worst scores for each individual respondent first and then calculate the sum across all the respondents in the sample. The other way is to calculate the sum of best counts (the most chosen items) and the sum of worst counts (the least chosen items) first and then calculates the difference between the two sums for each item (advertisement), as shown a hypothetical example in Table 1. Mathematically,it can be demonstrated that the two methods lead to the same results.

The best–worst scores can be standardised or normalised both at the individual level and at the aggregate level. At the individual level, one can derive the standard scores by dividing the best–worst scores for each individual respondent by r, where r is the numberof times each item appeared in the BWS task. This means that the standard scores for each individual will range from -1 to +1. Therefore, calculation of best–worst score is shown in below Equation: 3.

$$\text{Std. Score} = \frac{\text{Count}_{\text{Best}} - \text{Count}_{\text{Worst}}}{r \times n}$$

(Adapted from Messey et. al., 2013)

where$\text{Count}_{\text{Best}}$ = total number of times each item was judged 'best'; $\text{Count}_{\text{Worst}}$ = total number of times each item was judged 'worst'); n = number of respondents or sample size; and r = the frequency of theappearance of each item in the choice sets.

A hypothetical example has been shown below for the aggregate sample of 103 respondents.

Table 2: Calculating the Best-Worst Score

Rank	Advertisement	'Best' (most unethical ad)	'Worst' (least unethical ad)	Best−worst score	Standard score
1	Shoes	340	6	334	0.81
2	Cookies	225	49	176	0.43
3	Instant Noodles	177	102	75	0.18
4	Energy Drink	124	88	36	0.09
5	Jelly	156	169	−13	−0.03
6	Vitamins	76	141	−65	−0.16
7	Toothpaste	74	182	−108	−0.26
8	Antiseptic Soap	42	237	−195	−0.47
9	Milk	22	262	−240	−0.58
	Sum	1236	1236	0	0.00

(Adapted from Messey et al., 2013)

As can be seen from the BWS scores presented in the above hypothetical example, the results suggest that theshoe advertisement is perceived to be the most unethical as its standardised score is +0.81, this figure arrived by simply substituting the appropriate values into Equation 3. Similarly, calculation for all other advertisements, including the milk advertisement, which was rated the least unethical with a score of -0.58.

Generally, the BWS results confirm which of the item areperceived by the target audience as the most or least. Also, by using BWS wewere able to precisely quantify the levels of values and thus choose the correct item/values for use in the related study.

Limitations of Best- Worst Scaling Method

However, the extant literature remarked that this technique has some limitations. Firstly, this paper noticed from prior literature that it could be a complementary technique not as a complete substitute in doing research. Another limitation is that designing the choice set sometimes would not be accurate if respondents are indeed completely indifferent to a set of items being evaluated, e.g., if one offered a respondent a choice set with 3 items that were perceived to be equally bad or good, a BWS task would still elicit two choices. However, every method has its own superiority and drawbacks.

Conclusion

The Best-Worst scaling method is a simple procedure but powerful substitute to the most commonly used methods, many of which rely on researchers' own subjective judgments, or small sample qualitative research approaches of limited generalisability, or use rating scales which are subject to various types of response bias. No doubt, the BWS has wider applications in marketing and consumer behaviour literature as it allows researchers, advertisers, or policy-makers to choose what attributes/items should be evaluated and find the most and least important ones. BWS could also be applied to concept testing during new product development, to unambiguously identify concepts with the greatest commercial potential, and a follow-up BWS study could be used on the initial advertising campaigns with 'purchase intention' as the focal variable. In addition, BWS could be used to evaluate products in competitors' portfolios, and provide a powerful basis for comparing the attractiveness of different firms' offers to the market. Moreover, BWS can be used wherever researchers need valid, evidence-based, respondent-generated evaluations, e.g., of political platforms and political candidates. Political investigators using BWS would be able to more accurately identify the most important issues within electorates, and to identify the most popular, or electable candidate. Therefore, certainly, the Best-Worst scaling method shows a new horizon of knowledge to the academic researchers.

References

Auger, P., Devinney, T. M. and Louviere, J. J. (2007). Using best–worst scaling methodology to investigate consumer ethical beliefs across countries. *Journal of Business Ethics*, 70(3), 299-326.

Almquist, E. and J. Lee. 2009. "What Do Customers Really Want?" *Harvard Business Review* 84: 23.

Bednarz, A. (2006). Best-worst scaling and its relationship with multinomial logit. Bachelordissertation, University of South Australia.

Burke, P. F., Schuck, S., Aubusson, P., Buchanan, J., Louviere, J. J. and Prescott, A. (2013). Why do early career teachers choose to remain in the profession? The use of best–worst scaling to quantify key factors. *International Journal of Educational Research*, 62(0), 259-268. doi: http://dx.doi.org/10.1016/j.ijer.2013.05.001.

Chapman, R. G. and Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*, 19(3), 288-301.

Chrzan, K. and Golovashkina, N. (2006). An empirical test of six stated importance measures. *International Journal of Market Research*, 48(6), 717-740.

Cohen, S. (2003). Maximum difference scaling: improved measures of importance and preference for segmentation.Paper presented *at the Sawtooth Software Conference Proceedings*.

Cohen, S., and Neira, L. (2003).Measuring preference for product benefits across countries: overcoming scale usage bias with maximum difference scaling. Paper presented in the *ESOMAR 2003 Latin America Conference Proceedings*.

Cohen, S. H. and P. Markowitz (2002). "Renewing Market Segmentation: Some New Tools to Correct Old Problems." Paper presented at ESOMAR 2002 Conference, Amsterdam, The Netherlands.

Craig, C. S. and S. P. Douglas (2000). International Marketing Research. 2nd ed. New York: Wiley.

Finn, A., and Louviere, J. J. (1992). Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy & Marketing*, 11, 19-25.

Flynn, T. N., Louviere, J. J., Peters, T. J. and Coast, J. (2007). Best-worst scaling: What it can do for health care research and how to do it. *Journal of Health Economics*, 26(1), 171-189.

Green, P. E. (1974). On the design of choice experiments involving multifactor alternatives. *Journal of consumer research*, 1, 61-68.

Kahle, L. R. (1983). *Social values and social change: Adaptation to life in America*: Praeger Publishers.

Lancsar, E., Louviere, J. and Flynn, T. (2007). Several methods to investigate relative attribute impact in stated preference experiments. *Social Science & Medicine*, 64*(8)*, 1738-1753.

Lee, J. A., Soutar, G. and Louviere, J. (2008). The best–worst scaling approach: an alternative to Schwartz's values survey. *Journal of personality assessment*, 90(4), 335-347.

Lee, J. A., Soutar, G. N. and Louviere, J. (2007). Measuring values using best-worst scaling: The LOV example. *Psychology & Marketing*, 24(12), 1043-1058.

Louviere, J., Lings, I., Islam, T., Gudergan, S. and Flynn, T. (2013). An Introduction to the Application of (Case 1) Best-Worst Scaling in Marketing Research. *International journal of research in marketing*, 30, 292-303.

Louviere, J. J. and Islam, T. (2008). A comparison of importance weights and willingness-to-pay measures derived from choice-based conjoint, constant sum scales and best–worst scaling. *Journal of Business Research*, 61(9), 903-911.

Louviere, J. and Woodworth, G. (1990). Best-worst scaling: A model for the largest difference judgments: Working paper. University of Alberta.

Marley, A. A. J. and Louviere, J. J. (2005). Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology*, 49(6), 464-480. doi: http://dx.doi.org/10.1016/j.jmp.2005.05.003.

McFadden, D. (1986). "The Choice Theory Approach to Market Research." *Marketing Science* 5: 275–297.

Mullen, M. R. (1995). "Diagnosing Measurement Equivalence in Cross-national Research." *Journal of International Business Studies*, 26: 573–596.

McCarty, J. A. and L. J. Shrum (2000). "The Measurement of Personal Values in Survey Research: A Test of Alternative Rating Procedures." *Public Opinion Quarterly*, 64: 271–298.

Massey, G. R., Wang, P. Z., Waller, D. S. and Lanasier, E. V. (2013). Best–worst scaling: A new method for advertisement evaluation. *Journal of Marketing Communications*, 1-25. doi: 10.1080/13527266.2013.828769.

Raghavarao, D. and Padgett, L. V. (2005). *Block Designs: Analysis, Combinatorics and Applications* (Vol. 17). Hackensack, NJ: World Scientific.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology*, 25, 1-62.

Schwartz, S. H. and Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology*, 53(3), 550.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, 34(4), 273-286.

Appendix A: Steps to conduct a BWS study (Adapted from Messey et al., 2013)

1. See Appendix B and choose the appropriate BIBD for the number of items/attributes you wish to test. If the BIBD design is asymmetrical (i.e., k is not equal to s), you need to randomise the presentation order of the items in each choice set to control for possible order effects. No need for randomisation if the BIBD is symmetrical (e.g., 4, 5, and 7 items in the Appendix B).

2. Gather a representative sample of respondents and expose them to the items as per each choice set.

3. After each exposure to a choice set, record the BWS data, and do this until all choice sets have been shown.

4. Calculate the BWS scores using Equation 3 and place results in a table similar toTable 2.

5. Test for data reliability by ensuring that:

- The sum of the 'best' and 'worst' should be equal to each other.
- The sum of the 'best–worst score' column should be zero.
- The range for the Standard Scores should be between -1 and +1.

6. Interpret the results.

Appendix B: Sample balanced incomplete block designs (Adapted from Messey et al. 2013)

Note:

1. The BIBDs that are symmetrical are those where k = s.
2. In the below table, items/ attributes presented as Advertisement ID.

BIBD involving 4 ads/items

Set ID	Advertisement ID		
1	2	3	1
2	4	1	2
3	1	4	3
4	3	2	4

$k = 4, s = 4, r = 3, m = 3, \lambda = 2$

BIBD involving 5 ads/items

Set ID	Advertisement ID			
1	1	2	4	3
2	5	1	3	2
3	2	4	5	1
4	3	5	1	4
5	4	3	2	5

$k = 5, s = 5, r = 4, m = 4, \lambda = 3$

BIBD involving 6 ads/items

Set ID	Advertisement ID		
1	1	2	5
2	2	3	6
3	3	4	2
4	4	1	3
5	2	5	4
6	3	5	6
7	4	6	5
8	1	2	6
9	5	1	3
10	6	4	1

$k = 6, s = 10, r = 5, m = 3, \lambda = 2$

BIBD involving 7 ads/items

Set ID	Advertisement ID		
1	2	6	4
2	1	4	5
3	4	7	3
4	3	2	1
5	7	5	2
6	6	1	7
7	5	3	6

$k = 7, s = 7, r = 3, m = 3, \lambda = 1$

BIBD involving 8 ads/items

Set ID	Advertisement ID			
1	8	2	3	5
2	1	4	7	6
3	8	3	4	6
4	2	5	1	7
5	8	4	5	7
6	3	6	2	1
7	8	5	6	1
8	4	7	3	2
9	8	6	7	2
10	5	1	4	3
11	8	7	1	3
12	6	2	5	4
13	8	1	2	4
14	7	3	6	5

$k = 8, s = 14, r = 7, m = 4, \lambda = 3$

BIBD involving 9 ads/items

Set ID	Advertisement ID		
1	2	4	8
2	1	4	5
3	4	7	9
4	3	4	6
5	1	2	3
6	2	5	7
7	2	6	9
8	1	8	9
9	5	6	8
10	3	7	8
11	1	6	7
12	3	5	9

$k = 9$, $s = 12$, $r = 4$, $m = 3$, $\lambda = 1$

BIBD involving 10 ads/items

Set ID	Advertisement ID			
1	4	7	8	9
2	3	6	8	10
3	2	5	9	10
4	1	8	9	10
5	4	5	6	10
6	3	5	7	9
7	2	6	7	8
8	1	5	6	7
9	2	3	7	10
10	2	4	6	9
11	3	4	5	8
12	1	2	3	4
13	1	4	7	10
14	1	3	6	9
15	1	2	5	8

$k = 10$, $s = 15$, $r = 6$, $m = 4$, $\lambda = 2$

Technical Appendix: SAS syntax for generating balanced incomplete block designs

More complex BIBD designs beyond those listed in Appendix B can be constructed using computer programs such as SAS. The SAS/STAT® market research experimental design macros (Kuhfeld 2009) are powerful tools for constructing BIBD designs. An example of SAS syntax to construct a BIBD with $k = 10$, $s = 15$, $r = 6$, $m = 4$, $\lambda = 2$ is as follows:

```
%mktbibd    (b = 15,        /* 15 choice sets or blocks */
             t = 10,        /* 10 items or treatments */
             k = 4,         /* 4 items per choice set */
             seed = 350)    /* random number seed */
```

Note: SAS syntax uses a notation that is different from that used in this paper. In SAS notation, $b = s$, $t = k$, $k = m$.