How Good Is Best-Worst Scaling?

And Is It More than Just Rank-Order Data?

Bryan Orme, Sawtooth Software[1]

March 2018

Executive Summary

Best-worst scaling (BWS) gives you better information with fewer respondents—it works better than traditional rating scales or constant sum questions, achieving better discrimination among the items and finding a larger number of statistically significant differences between groups of respondents. BWS yields more than just rank-order scaling and is able to recover relative metric differences among items. BWS provides more accurate information than 10-point ratings and constant sums for classifying respondents and making individual-level predictions. If the BWS scores need to be scaled relative to a buy/no buy or important/not important threshold, anchored BWS is a straightforward solution.

Introduction

Researchers often are asked to measure the preference or importance of items such as product features, claims, packaging styles, or health risks. Popular approaches include rating scales, constant-sum tasks, and best-worst scaling (MaxDiff).

The standard 5- and 10-point rating scales are fast and easy, but are plagued by low discrimination among items and scale-use bias. With rating scales, a much more important item won't necessarily get a much larger score. Respondents are tempted to yea-say (positivity bias) and straight-line. Different cultural backgrounds can influence the way respondents use the scale. When comparing groups of respondents, many true differences may be obscured due to the messiness and bias in rating scale data.

With constant-sum tasks, respondents are asked to distribute (say) 100 points across multiple items. Many respondents struggle with this task and the results can be noisy and imprecise. Constant-sum tasks are also difficult for respondents to do with more than about ten items.

Best-worst scaling (BWS) shows sets of items, typically four or five at a time (Exhibit 1), asking for each set which item is the best and worst (or most and least important) (Louviere 1991, Finn and Louviere 1992, Louviere et al. 2015). It is typical to show each respondent eight to fifteen BWS sets, such that each item appears at least once and preferably two or three times per respondent in a balanced design. Often 20 or more items are included in a BWS study.

[1] The author thanks David Lyon, Keith Chrzan, and Tom Eagle for their critique on earlier drafts of this article. This article originally published in *Quirk's* (July 2018).

Think about what makes you choose one restaurant or another. Considering only these four features, which is the Most Important and which is the Least Important?

(5 of 8)

Most Important		Least Important
○	Restaurant donates generously to charities	○
○	Food tastes wonderful	○
○	Restaurant is close to your home	○
○	Clean eating area	○

Exhibit 1

BWS scores show greater differences among the items and the results are more predictively accurate of held out information (Cohen & Orme 2004, Chrzan and Golovshkina 2006). You will find a greater number of statistically significant differences among the items and between the respondents (Cohen & Orme 2004). As a result, with BWS you can use smaller sample sizes and obtain equally good estimates as the competing methods.

The main drawback for BWS is that it can take about triple the time for respondents to do than rating scales. But given the way respondents often rush through ratings grids and tend to yea-say and straight-line, having respondents slow down and provide better data seems like a good thing for the conscientious researcher to do.

Does Best-Worst Provide More than Rank Data?

We have seen false claims that best-worst scaling (BWS) data yield nothing more than rank-order information. The argument is as follows: assume respondents use BWS to evaluate items that have an established preference order. In each set, if respondents pick the true best and worst items without error there is no way to learn that there might be a much greater difference in metric preference score between the first and second vs. the second and third items, etc.

The fallacy in the above narrative is the false assertion that respondents answer like robots, mechanically selecting the true best and worst items in each BWS set. Researchers typically study complicated constructs such as importance or preference and humans don't have a fixed list of scores that they somehow pull out of their brains to reference as they answer BWS questions. Their answers are subject to inconsistencies (errors). Imagine objects A and B are clearly different for a respondent, but objects C and D are extremely close. It's likely that each time the respondent compares A and B in a set, she picks A over B. But, that same respondent might not be consistent over multiple sets regarding the order of C and D because of how nearly identical they are in her mind.

Using robotic respondents programmed to act like imperfect humans (let's call them artificial respondents), we can demonstrate that BWS can capture more than just rank-order scaled preferences: it can accurately recover the relative metric differences among the items. Consider a 9-item BWS study where the true utilities follow non-equal step sizes (Table 1).

Utilities for Artificial Respondents		
Item #	True Utilities	N=750 Estimated
1	5.00	5.03
2	4.10	4.14
3	3.90	3.91
4	3.00	3.03
5	2.10	2.09
6	1.90	1.90
7	1.50	1.60
8	1.00	1.07
9	0.00	0.00

Table 1

We generated a BWS questionnaire and programmed human-imitating bots to answer according to the true utilities in Table 1, subject to the typical level of response error expected from humans with distributional properties consistent with logit estimation. In Table 1, we also show the estimated scores for these 750 artificial respondents and they very closely fit the irregular utility intervals between adjacent items in this study. BWS has yielded metric information beyond just rank-order scaling. Given enough artificial respondent bots, we can make the estimated utilities perfectly match the true utilities to as many decimal places of precision as desired.

But can we demonstrate that BWS captures relative metric information with real respondents? We recently conducted a BWS study using 350 respondents from Amazon's Mechanical Turk panel sample. The challenge was to select items with a known quantity metric so that there was an absolute truth to validate against, so we decided to use the population of nine European countries. As a training task, we first showed respondents the nine countries with their populations (in alphabetical order). Following the training task, we asked respondents a few tangential questions regarding which of the nine countries they had visited before and which countries would be best to visit in summer or winter.

Next came the main focus of the study. We asked respondents to rate the nine countries based on relative population size using a 10-point rating scale (endpoints labeled 1= "Less population," 10="More Population"), a 100-point constant sum scale, and using BWS (7 sets of 4 items each). Respondents completed all three exercises and we randomized the order to control for order bias. In the BWS exercise (labeled "most population," "least population"), respondents saw each country at least three times. We averaged the scores across respondents for the three scaling approaches (Table 2).

European Country	True Population Size (Millions)	Best-Worst Scaling Results N=350	Constant Sum Results N=350	1-10 Point Rating Scale Results N=350
Germany	82.29	25.43	18.22	8.47
France	65.23	18.60	14.30	7.72
United Kingdom	65.11	17.73	14.47	7.77
Italy	59.32	11.98	11.28	6.70
Spain	46.40	9.18	10.65	6.32
Ukraine	44.01	5.72	8.75	5.25
Poland	38.10	5.53	8.36	5.18
Romania	19.62	3.07	6.96	4.43
The Netherlands	17.08	2.75	7.00	4.18
Sum:	437.16	100.00	100.00	56.02

Table 2

The rank-order recovery for the nine countries' populations was perfect for BWS and nearly perfect for 1-10 and constant sum scales. The next thing we noticed is that, true to previous research, BWS led to larger ratio differences among the scores compared to constant-sum and especially 10-point rating scales. The degree of response error affects how much more accentuated the BWS scores become, but in every case we've compared to the 1-10 rating scale, BWS shows greater differentiation. (We used an exponential transform of the logit estimated utilities to place the BWS results on a ratio scale with a meaningful zero point.)

The largest four countries offered the opportunity to see whether BWS could capture relative metric differences beyond just rank-order information (Exhibit 2).

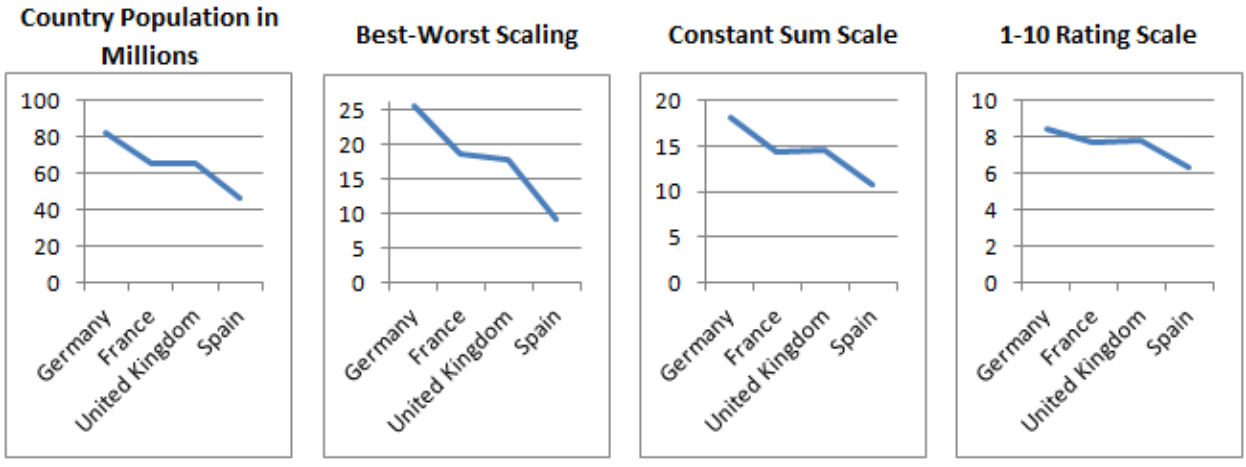

Exhibit 2

Remember, we reminded respondents regarding the true country populations at the beginning of the questionnaire. France (65.23 million) and the UK (65.11 million) have nearly the same population. Germany is larger with 82 million and Spain smaller with 46 million people. How well did the three methods recover the fact that France and the UK essentially have the same population?

Referring to Exhibit 2, BWS appears to be capturing something more than just rank order: it recovers the fact that France and the UK's populations are much closer relative to Germany and Spain, along with directionally showing that France has slightly larger population than the UK.

Looking just at average scores across a sample can cover up a lot of ugliness happening at the individual level. Those individual-level problems can be especially detrimental when classifying, clustering, or segmenting respondents and making predictions. As a straightforward individual-level consistency test, we examined whether the average of the scores for the true top 3 countries, middle 3, and bottom 3 were in descending order. If the three summary estimates were in the correct order for a respondent we scored a hit for the respondent, otherwise we scored a miss. BWS (via HB estimation) showed 80% individual-level consistency, the 10-point rating scale had 65%, and the constant-sum scaling showed 62% consistency. HB's ability to "borrow" data across the sample probably is cheating for this individual-level validation, so we re-estimated the BWS utilities without any data borrowing via purely individual-level logit estimation and found that the BWS consistency score was still superior at 72% ($p<0.01$).

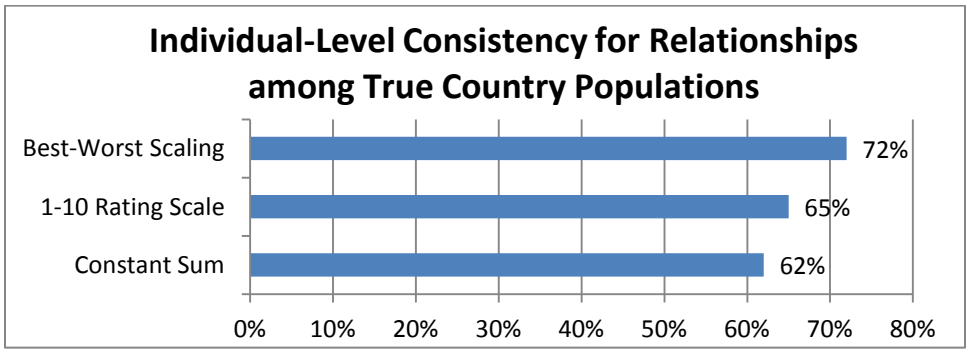

Exhibit 3

This advantage in terms of individual-level consistency (Exhibit 3) helps explain BWS's superiority for classification, clustering, and prediction relative to the rating scale methods.

We recognize that our experiment focused on the ability of respondents to recall factual information rather than their ability to express preference or importance. There are psychological differences, for certain. Given that we have already conducted and seen multiple comparative studies among these methods on preference/importance scales, we thought it would add to the discussion to conduct this research involving memory recall of objects with known and true metric measures.

But Aren't Best-Worst Scores Just Relative?

Best-worst scaling (BWS) involves just comparative judgments. We don't learn whether respondents like or dislike the items in any absolute sense. One solution is to include an item in your study that has some known preference, monetary value, or importance. You can then compare the score for the known item to the other items in the experiment. Another solution is *anchored BWS*: adding a few questions to a BWS study to establish a meaningful anchor point, such as buy/no buy, like/dislike, or important/not important.

We applied anchored BWS to the evaluation of European country sizes. After we asked the BWS questions, we computed the scores for each individual, on-the-fly. We ranked the countries from first to ninth based on each respondent's scores and asked respondents whether certain countries had more or

less than 50 million people. Rather than ask for all nine countries, we shortened the task by asking about only the 1st, 3rd, 6th, and 9th place countries in terms of the respondent's individualized ordering.

After re-estimating the model with the addition of the anchoring information, the new BWS scores now were scaled relative to an anchoring point of 50 million in population, with the anchor located correctly between Italy (59 million) and Spain (46 million people).

Conclusion

With or without the optional anchoring step, best-worst scaling (BWS) yields more than just rankings. Using both robotic and real respondents, we've shown that it recovers relative metric differences among items. If obtaining discriminating measurement of preference and importance is critical to you, if you want to avoid scale use bias, and if you want to improve your market segmentation and predictions, then BWS is for you. It is a truly an exceptional tool in the researcher's toolbox.

References

Chrzan, Keith and Natalia Golovashkina (2006) "An Empirical Test of Six Stated Importance Measures," *International Journal of Market Research*, 48, 717-40.

Cohen, Steve and Bryan Orme (2004), "What's Your Preference?" *Marketing Research*, Summer, American Marketing Association.

Finn, A. and J. J. Louviere (1992), "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety," *Journal of Public Policy and Marketing*, 11, 1, 12-25.

Louviere, Jordan J. (1991) "Best-Worst Scaling: A Model for the Largest Difference Judgments." Working paper. University of Alberta.

Louviere, Jordan J., Terry N. Flynn, and A. A. J. Marley (2015), "Best-Worst Scaling: Theory, Methods and Applications," Cambridge University Press.