Threshold Determination Using Extensions of Best-Worst Scaling

Sven Beisecker
WHU - Otto Beisheim School of Management
Christian Schlereth
WHU - Otto Beisheim School of Management

Cite as:
Beisecker Sven, Schlereth Christian (2020), Threshold Determination Using Extensions of Best-Worst Scaling. *Proceedings of the European Marketing Academy*, 49th, (59336)

Paper from the 49th Annual EMAC Conference, Budapest, May 26-29, 2020.

Threshold Determination Using Extensions of Best-Worst Scaling

Abstract:

Best-worst scaling (BWS) is a popular method that seeks to measure preferences for a large number of items (typically more than 7) between two extremes of a continuous scale (e.g., “best” and “worst”). However, BWS suffers from the problem that identifying a threshold is not possible. This means that a researcher cannot distinguish which of the items actually belong to the “best” or “worst” category. The present study (i) proposes and compares preference measurement approaches that include an indirect threshold question in BWS and (ii) develops a method that uses this information to estimate the threshold. In an exemplary study on people’s willingness to engage in climate-protecting actions, the approaches are compared in terms of choice consistency, response time, cognitive ease of survey completion, and the distribution of yielded thresholds. We thereby show that eliciting less rather than more information through the threshold question can be advantageous.

Keywords: best-worst scaling, threshold, climate-protecting actions

Track: Methods, Modelling & Marketing Analytics

1 Introduction

The best-worst scaling method is an extension of the traditional paired comparison method to multiple choices that asks participants in a sequence of comparison sets to choose the best and the worst item of a collection. The terms “best” and “worst” simply constitute a metaphor for two extremes of a latent, subjective continuum (Dyachenko, Reczek, & Allenby, 2014; Louviere, Lings, Islam, Gudergan, & Flynn, 2013). With this research, we aim to address BWS’ problem that identifying a threshold to distinguish whether an item belongs to the “best” or “worst” category is not possible. As an example, assume a study which aims to assess elements that might be important for an excellent restaurant experience and that these elements are described by the items “food quality”, “friendliness of the waitress”, “cleanliness”, “interior design”, “good music”, and so on. Using best-worst scaling, an analyst obtains a preference ranking of these items, however, he or she cannot infer if one of the low-ranked components could ultimately destroy the restaurant experience. Let’s assume that “interior design” takes on the middle position and “good music” is ranked directly below. In this case, it is impossible to infer whether the music will ultimately deter that customer from future visits.

One way to elicit additional information lies in the extension of traditional BWS by an indirect threshold question. Herein, respondents are asked to make an absolute judgment on either a subset or all of the items in a choice set in addition to the conventional best-worst choice task. Louviere, Flynn, and Marley (2015) suggest such a dual-response option in which respondents are asked to indicate, after completing a best-worst choice task, whether “all”, “some”, or “none” of the objects in a choice set are “good”.

However, no academic insights have been gained regarding the determination of an absolute threshold thus far. Against this background, the aim of the present study is (i) to propose and compare preference measurement approaches that include an indirect threshold question in BWS and (ii) to develop a method that uses this information to estimate a threshold.

In section 2, we motivate two operationalizations of the dual-response option in addition to the one suggested by Louviere et al. (2015) and describe a method to determine an absolute threshold. In section 3, we introduce an empirical study in which the different survey versions are compared across metrics such as choice consistency, response time, cognitive ease of survey completion, and the distribution of yielded thresholds. Section 4 offers concluding remarks.

2 Methodology

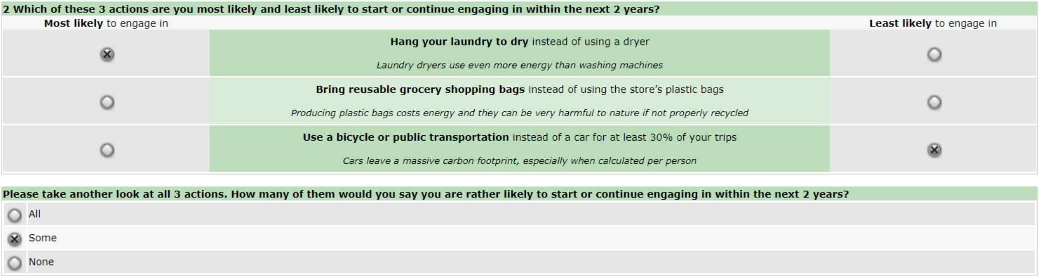
2.1 Problem of identifying a threshold

We illustrate the underlying threshold problem on the popular and simple count method to derive best-worst scores (e.g., Kaufmann, Rottenburger, Carter, & Schlereth, 2018): Assuming a balanced and orthogonal design for the experiment, we add a score of 1 for every time a given item has been picked as best, and subtract a score of 1 for every time the item has been picked as worst (Louviere et al., 2015). Consequently, the best-worst scores range between [-number of repetitions of an item; +number of repetitions of an item]. Some of the best-worst scores might equal zero. However, this zero is not suitable to distinguish whether a given item actually belongs to the category “best” or “worst”.

The reason is that being assessed with a best-worst score of zero simply depends on the composition of the items that have been chosen upfront by the researcher for the study. The only case in which the zero can distinguish “best” items from “worst” items would be if a researcher already had a priori knowledge about the categorization of the items and had picked the same number of items that certainly belong to the category of “best” and “worst”. However, such a scenario is unlikely. In reality, it is, for instance, also possible that all but one item belongs to the category of “best”. The resulting problem still holds when using more sophisticated methods for the estimation, such as the logit model (Hinz, Schlereth, & Zhou, 2015; Marley & Louviere, 2005), because these results are just proportional to the best-worst scores.

2.2 Threshold measurement

Assume a best-worst study in which three items are assessed in each choice set. In order to elicit further information, the threshold question proposed by Louviere et al. (2015) is asked in addition to each choice set, requesting respondents to indicate whether they perceive “all”, “some”, or “none” of the items as good (V1). An exemplary choice set is displayed in Figure 1.



2 Which of these 3 actions are you most likely and least likely to start or continue engaging in within the next 2 years?

Most likely to engage in		**Least likely** to engage in
☒	**Hang your laundry to dry** instead of using a dryer *Laundry dryers use even more energy than washing machines*	○
○	**Bring reusable grocery shopping bags** instead of using the store's plastic bags *Producing plastic bags costs energy and they can be very harmful to nature if not properly recycled*	○
○	**Use a bicycle or public transportation** instead of a car for at least 30% of your trips *Cars leave a massive carbon footprint, especially when calculated per person*	☒

Please take another look at all 3 actions. How many of them would you say you are rather likely to start or continue engaging in within the next 2 years?

○ All

☒ Some

○ None

Figure 1. Exemplary Choice Set (V1)

We believe that the all-some-none question bears certain disadvantages: According to the compromise effect (Simonson, 1989), there may be a high tendency for respondents to pick the "some" option. If true, this tendency would limit the amount of additional information that can be gained. Second, the answer option "some" is ambiguous for the middle item, i.e., the one that has neither been evaluated as "best" nor as "worst". When a respondent picks "some", it is still unclear for this item whether it rather belongs to the category of "good" or "bad" items.

In this paper, we aim to overcome these shortcomings by testing two modifications: Version 2 (V2) asks explicitly for the number of items in a choice set that are considered "good" and "bad", respectively. Respondents need to make an evaluation of all items. As such, we obtain information for the middle item and avoid the compromise effect. Version 3 (V3) also accomplishes that goal, however, asks respondents to make a judgment only on the item which they neither evaluated as "best" nor as "worst". Table 1 summarizes the three versions in terms of their respective response options.

V1 (Louviere)	*V2 (new)*	*V3 (new)*
All items are good	All 3 items are good (3-0)	Middle item is good
Some items are good	2 items are good, 1 item is bad (2-1)	Middle item is bad
None of the items are good	1 item is good, 2 items are bad (1-2)	
	All 3 items are bad (0-3)	

Table 1. Three Versions of the Threshold Question

2.3 *Estimation*

Having determined the best-worst scores according to Louviere et al. (2015), we devise a method to determine a threshold based on which items can be attributed an absolute value. Assume each item occurs four times across choice sets so that best-worst scores range from -4 to +4. In this setup, any decision to assign a certain best-worst score, e.g. zero, the role of a threshold would be arbitrary. However, using the response from the additional question, we can shift the scale such that a best-worst score of zero on a new scale becomes a meaningful threshold. Figure 2 illustrates this idea.

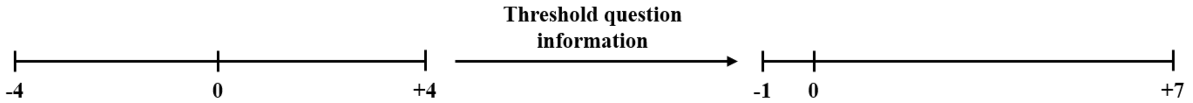


Figure 2. Illustration of Scale Shift

In the following, we describe the mathematical underpinnings, exemplary for V3: Given the best-worst score BW$_{h,i}$ of an item i and respondent h, the probability of that item being evaluated as "good" in the threshold question is given by:

$$\Pr(x_{h,i}=1)=\frac{e^{\beta_{h,0}+\beta_{h,1}*BW_{h,i}}}{1+e^{\beta_{h,0}+\beta_{h,1}*BW_{h,i}}}\ \left(i\in I,h\in H\right) \tag{1}$$

Subsequently, the objective is to find a set of parameters $\beta_{h,0}$ and $\beta_{h,1}$ such that the choice probabilities implied by the best-worst score of an item are fitted as close as possible to the threshold decisions. Summing across all choice sets, we maximize the following objective function for the case of V3:

$$\max_{\beta_{h,0},\beta_{h,1}}\sum_{j\in C_j}\ln\left(\Pr\left(x_{h,m,j}=1\right)\right)*t_{h,j}+\ln\left(1-\Pr\left(x_{h,m,j}=1\right)\right)*\left(1-t_{h,j}\right)\ (h\in H) \qquad (2),$$

where j represents the choice set index, t$_{h,j}$ the corresponding threshold decision, and x$_{h,m,j}$ the middle item in choice set j. Based on a maximum likelihood estimation (here, Excel Solver is already sufficient), we find the set of parameters $(\beta_{h,0},\beta_{h,1})$ that fits the best-worst-implied choice probabilities as close as possible to the respective threshold decisions.

Finally, we intend to find the best-worst score which coincides with the threshold. To model choice indifference, the probability from equation 1 is set equal to .5. Solving for BW, we obtain the score that, on the current scale, coincides with the threshold:

$$\beta_{h,0}+\beta_{h,1}*BW_{h,\ Threshold}\overset{!}{=}0\Rightarrow BW_{h,\ Threshold}=-\frac{\beta_{h,0}}{\beta_{h,1}}\ (h\in H) \tag{3}$$

We can now shift the best-worst scale such that the determined threshold is assigned a new best-worst score of zero. Accordingly, each item obtains a shifted best-worst score that lies on this new scale via the following equation:

$$BW_{h,i}-(-\frac{\beta_{h,0}}{\beta_{h,1}})\ \left(i\in I,h\in H\right) \tag{4}$$

Items with a positive [negative] shifted best-worst score can now be categorized as "good" ["bad"], respectively.

We note that the estimations for V1 and V2 are similar: V1 and V2 have comparable objective functions, with the difference that not only the middle item, but also the best and worst item in each choice set must be considered.

3 Empirical Study

3.1 Setup

In an exemplary study, respondents were asked to express their opinion on different environment-protecting actions. We employed a balanced incomplete block design (BIBD) with twelve choice sets of three items each (Louviere et al., 2015; Louviere et al., 2013). Items were sampled from a set of nine actions, which are shown in Table 2. In each of the choice sets, respondents were asked to select the action they would "most likely" and "least likely" start or continue to engage in within the next two years.

	Action
1	Use a bicycle or public transportation instead of a car for at least 30% of your trips
2	Bring reusable grocery shopping bags instead of using the store's plastic bags
3	Use an electricity provider that provides you only energy from renewable sources
4	Eat less meat, specifically less than three times per week
5	Bring a reusable coffee mug instead of using paper or plastic cups, for instance in cafés or at work
6	Hang your laundry to dry instead of using a dryer
7	Only buy glass bottles instead of plastic bottles
8	Turn off electronic devices and equipment when not using them instead of leaving them in stand-by mode
9	Buy regional products rather than foreign ones for at least 50% of your groceries, even though they might be more expensive

Table 2. List of Environment-Protecting Actions Presented in Best-Worst Study

Subjects were matched to one of four versions of the same survey at random. A sample size of 404 [V1: 97, V2: 104, V3: 105, V4: 98] was obtained. The survey versions did not differ in terms of the best-worst task, however, in terms of the threshold question. The threshold questions of V1 to V3 coincide with those presented in section 2.2. Only the wording is adjusted to the topic at hand, asking respondents in this case to make the absolute decision whether they are "rather likely" or "rather unlikely" to engage in the respective actions. In addition, a fourth version (V4) was included which represents a conventional best-worst task without extensions and serves to test whether respondents behave differently in the best-worst tasks when the indirect threshold question is not included.

3.2 Comparison of survey versions

In a first step, we determined the set of best-worst scores for each respondent. As described in section 2.1, best-worst scores were computed by adding a score of 1 for every time a certain item has been picked as “most likely” and subtracting a score of 1 for every time the item has been picked as “least likely”. Based on the best-worst scores, a consistency score for each respondent was obtained. It is measured by the sum of squared best-worst scores across all items (Louviere et al., 2015, p. 29).

In the present study, a respondent provided perfectly consistent answers if each integer best-worst score in the range from -4 to +4 occurred exactly once across the nine items. The sum of squared best-worst scores takes on its maximum value of 60 in this case. In Table 3, we report the average consistency score across respondents in the four survey conditions. Conducting an ANOVA, no significant differences among consistency scores are detected across the four versions (*F*(3, 400) = 0.14, p = .939). Hence, the presence and type of threshold question do not have an impact on respondents’ consistency in choosing the actions they would “most likely” and “least likely” engage in.

Furthermore, respondents’ consistency in answering the threshold question was determined. As a general rule, an inconsistent decision results if the same item is declared to be “rather likely” in one choice set and “rather unlikely” in another. On a respondent level, consistency is measured as the percentage of items that, out of the presented nine, were consistently classified. In Table 3, the average consistency across respondents is reported for each survey version. Conducting an ANOVA, we note that there are significant differences in respondents’ threshold consistency among V1 to V3 (*F*(2, 303) = 23.60, *p*<.001). A Tukey’s honestly significant difference (HSD) test reveals that V2 yields a significantly lower threshold consistency than V1 and V3 do (*p*<.001), respectively. We also list the respective percentage of respondents that were consistent in their threshold decisions across all items. V3 performs best on this metric.

Apart from consistency, we measured how demanding the respective survey versions were to respondents. As a first measure, we considered the mean time it took respondents to complete the BWS section.[1] An ANOVA reveals significant differences in response time among survey versions (*F*(3, 383) = 4.15, *p*=.007). A Tukey’s HSD test shows weakly significant differences between V4 and V1 (*p*=.058), and significant differences between V4 and V2 (*p*=.014) as well as V4 and V3 (*p*=.016). Unsurprisingly, it took respondents the least time to complete V4 since they did not have to comprehend nor respond to a threshold

[1] We removed respondents with a substantial completion time gap (>1,000 seconds).

question. Among versions with a threshold question, however, no significant completion time differences exist.

The reported differences in mean completion time among survey versions diminish when excluding the first three choice sets from the comparison. Conducting an ANOVA, only weakly significant differences among survey versions are detected (*F*(3, 383) = 2.33, *p*=.074). According to Tukey's HSD test, the only significant difference is between V2 and V4 (*p*=.071). Hence, no significant differences in response time for the last nine choice sets can be detected among V1, V3, and V4, even though the former two versions contain a threshold question, while the latter does not. We conclude that an initial effort to understand the threshold question is the main driver of response time differences. Over time, a learning effect sets in and respondents internalize the nature of the respective threshold question to the point that they can answer the additional question without taking significantly more time.

Finally, we measured the cognitive ease of completing the different survey versions, using a four-item scale, which we adapted from Bettmann, John, and Scott (1986, p. 319). Fit indices corresponding to a one-factor model, as reported below Table 3, allow for the conclusion that a unidimensional model fits the data. Composite reliability is acceptable. An ANOVA reveals that there are no significant differences in terms of mean cognitive ease among the four survey versions (*F*(3, 400) = 1.68, *p*=.171). In other words, the presence and type of threshold question do not have a significant impact on perceived complexity.

	V1	*V2*	*V3*	*V4*
Consistency Index: $\frac{1}{H}\sum_{h=1}^{H}\sum_{i=1}^{9} BW_{hi}^{2}$	56.74	56.37	56.30	56.65
Consistent Threshold Choices	84.65%	73.72%	89.42%	-
Perfectly Consistent Respondents (Threshold)	26.80%	18.27%	52.38%	-
Mean Completion Time BWS Section [s]	335	345	345	281
Mean Completion Time Choice Sets 4-12 [s]	181	189	185	159
Mean Cognitive Ease Score[a]	5.06	5.26	5.04	5.39

[a] Composite of items *simple, easy, easy to follow, not difficult to complete*; items measured on a scale from 1 to 7; test of exact fit: *p*=.105, RMSEA=.056, TLI=.981, CFI=.994, SRMR=.016; CR=.771

Table 3. Comparison of Best-Worst Survey Versions

3.3 Threshold results

Following the methodology described in section 2.3, a threshold was determined for each individual and a corresponding scale shift conducted. The scale shift was truncated to conform to a minimum [maximum] value of -4 [+4]. Thereby, the entire spectrum of possible outcomes was covered, including the extreme cases of all items being "rather likely" (shift of -4) and of all items being "rather unlikely" (shift of +4), while containing the magnitude of the shift.

Shifting the scales of each respondent, we obtained histograms for V1 to V3 (Figure 3) which plot the ranks of the last item classified as “rather likely” against their relative frequency of occurrence across respondents. Intuitively, the rank represents the number of items that lie above the threshold. Ranks range from 0 (all items are classified as “rather unlikely”) to 9 (all items are classified as “rather likely”). The distribution shows that V1 is particularly prone to producing thresholds in extreme positions.

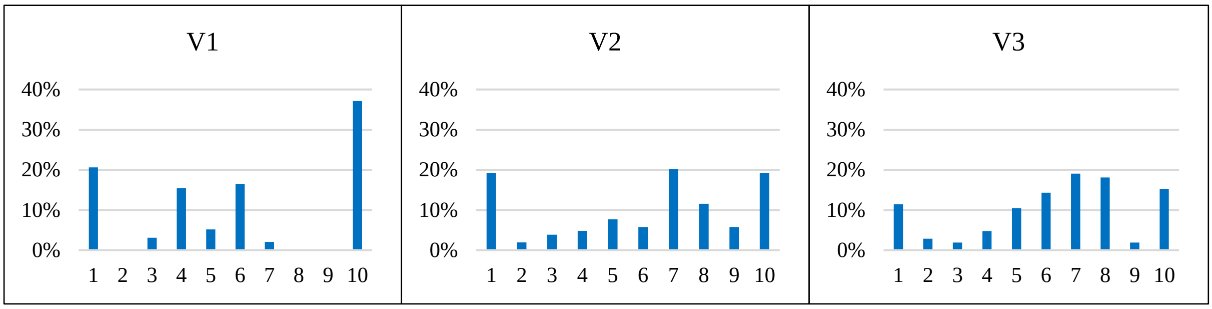


Figure 3. Distribution of Ranks

Finally, the shifted best-worst scores were averaged across respondents to obtain a global ranking of the survey-constituent items. The item rankings produced by V1 to V4 are summarized in Table 4. The ranking for V4 relies on the average of conventional best-worst scores since no threshold question was used. It serves as a control condition against which the rankings established by the other versions can be compared. We note that V3 produces the ranking that is most in line with V4, with a rank correlation of .983.

	Items			
Rank	*V1*	*V2*	*V3*	*V4*
1	2	2	2	2
2	6	6	6	6
3	9	1	1	5
4	5	5	5	1
5	1	9	3	3
6	8	8	9	9
7	7	4	8	8
8	3	7	4	4
9	4	3	7	7

Note: Rank correlations of V4 with V1 [V2, V3]: .783 [.817, .983]

Table 4. Item Rankings

4 Conclusion

In the present study, three different implementations of the threshold question for BWS were proposed and compared across several criteria. We demonstrated how a shift of the best-worst scale provides a meaningful zero point for each individual which separates “good” items from “bad” items, or, in this study, those actions respondents are “rather likely” from those they are “rather unlikely” to engage in.

Current results are in favor of V3. In specific, we learn that V1 and V3 yield a significantly higher consistency in answering the threshold question than V2 does. Similarly, V3 has the highest rate of perfectly consistent respondents, that is, those who were consistent in their threshold decisions across all items. V3 also produces the global ranking that is most aligned with V4, the baseline best-worst survey. The results speak for the fact that eliciting less (V3) rather than more (V2) information can, in fact, be advantageous. On a content level, the study sheds light on individuals' relative and absolute willingness to engage in different climate-protecting actions. Most notably, bringing reusable grocery shopping bags to stores leads the ranking across survey versions.

While the present study proposes a method to determine an absolute threshold from BWS by means of adding a threshold question, an external reference value against which to judge the accuracy of the threshold is missing. Therefore, future research should investigate ways to establish external validity with regard to threshold determination. Finally, variations of the threshold question other than the ones presented here could be tested. Possibly, more conclusive results regarding the favorability of different survey operationalizations could thus be obtained.

5 Literature

Bettmann, J. R., John, D. R., & Scott, C. A. (1986). Covariation assessment by consumers. *Journal of Consumer Research, 13*(3), 316-326.

Dyachenko, T., Reczek, R. W., & Allenby, G. M. (2014). Models of sequential evaluation in best-worst choice tasks. *Marketing Science, 33*(6), 828-848.

Hinz, O., Schlereth, C., & Zhou, W. (2015). Fostering the adoption of electric vehicles by providing complementary mobility services: A two-step approach using best–worst scaling and dual response. *Journal of Business Economics, 85*(8), 921-951.

Kaufmann, L., Rottenburger, J., Carter, C. R., & Schlereth, C. (2018). Bluffs, lies, and consequences: A reconceptualization of bluffing in buyer-supplier negotiations. *Journal of Supply Chain Management, 54*(2), 49-70.

Louviere, J., Flynn, T., & Marley, A. (2015). *Best-worst scaling: Theory, methods and applications*: Cambridge University Press.

Louviere, J., Lings, I., Islam, T., Gudergan, S., & Flynn, T. (2013). An introduction to the application of (case 1) best–worst scaling in marketing research. *International Journal of Research in Marketing, 30*(3), 292-303.

Marley, A. A. J., & Louviere, J. J. (2005). Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology, 49*(6), 464-480.

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research, 16*(2), 158-174.