Threshold Determination Using Extensions of Best-Worst Scaling

Sven Beisecker
WHU - Otto Beisheim School of Management
Christian Schlereth
WHU - Otto Beisheim School of Management
Felix Eggers
University of Groningen

Cite as:
Beisecker Sven, Schlereth Christian, Eggers Felix (2021), Threshold Determination Using Extensions of Best-Worst Scaling. *Proceedings of the European Marketing Academy*, 50th, (93642)

Paper from the 50th Annual EMAC Conference, Madrid, May 25-28, 2021

Threshold Determination Using Extensions of Best-Worst Scaling

Abstract:

Best-worst scaling (BWS) is a popular method that seeks to measure preferences for multiple items on a continuous scale between two extremes (e.g., “best” and “worst”). Yet, BWS suffers from the threshold identification problem, i.e., the obtained scores and rankings provide insights into each item’s relative preferences, but not into the overall acceptability of an item. For example, firms applying BWS to score different slogans will not know which of these, if any, are acceptable. The present paper (i) proposes different threshold identification approaches and (ii) develops models for the corresponding multinomial Hierarchical Bayes estimation. In two empirical studies we compare the approaches’ choice consistency, response time, cognitive ease of survey completion, and resulting parameter estimates. Although simulation results seem to advocate the elicitation of more information, empirical evidence shows that the simplest indirect threshold identification approach is on par.

Keywords: best-worst scaling, threshold, Hierarchical Bayes

Track: Methods, Modelling & Marketing Analytics

1 Introduction

The best-worst scaling (BWS) method is a rather young variant of traditional discrete choice experiments, which recently gained increasing popularity (Brynjolfsson, Collis, & Eggers, 2019; Dyachenko, Reczek, & Allenby, 2014; Louviere, Lings, Islam, Gudergan, & Flynn, 2013). It aims to measure preferences for multiple items by asking participants in a sequence of comparison sets to choose the best and worst options of a subset of the items. The terms "best" and "worst" constitute a metaphor for two extremes of a latent, subjective continuum. However, the estimated preferences are relative in nature, i.e., interval scaled, and cannot separate good from bad items.

With this research, we address BWS' threshold identification problem and propose approaches that aim to distinguish whether an item belongs to the "best" or "worst" category. For example, consider an automotive study that examines car brands' appeal, such as Tesla, Ford, BMW, and Dacia. Using BWS, an analyst obtains the individual ranking of a respondent's preferred cars. Still, how many of the brands and which are in the respondent's consideration set remains unknown.

The present paper (i) proposes different threshold identification approaches in BWS and (ii) develops models for the corresponding multinomial Hierarchical Bayes estimation. In two empirical studies we compare the approaches' choice consistency, response time, cognitive ease of survey completion, and resulting parameter estimates. We are able to show that the simplest indirect threshold identification approach is on par.

2 Literature Review

Since we are unaware of research that examines approaches for BWS' threshold identification problem, we first summarize research on traditional discrete choice experiments (DCEs) which illustrates that the way in which threshold identification questions are included into the choice task substantially affects the measured threshold location (Schlereth & Skiera, 2017).

2.1 Threshold identification in discrete choice experiments

Most frequently, researchers use a no-choice option that is presented as an additional alternative within each choice set. Dhar (1997) motivates the use of the no-choice option to allow for choice deferral and tests under which conditions respondents make use of this option. The no-choice option is found to provide a gateway for respondents who perceive the set of alternatives in a choice set as overall unattractive or prefer to continue the search until finding an alternative that exceeds their reservation utility (Dhar, 1997). Dhar and Simonson (2003) find evidence that the no-choice option leads to a decrease in the compromise effect and an increase in the attraction effect in choice tasks.

Including a no-choice option, however, comes at the cost of information loss since this option provides no insights into respondents' relative preferences and thus limits the basis for parameter estimation. This issue can be addressed by a dual response design consisting of a choice among a set of alternatives (forced choice) and a subsequent choice among that set of alternatives *and* the no-choice option (free choice) (Brazell et al., 2006; Dhar & Simonson, 2003). Further, Schlereth and Skiera (2017) propose the Separated Adaptive Dual Response (SADR) approach, which strictly separates all forced choice from all free choice questions and introduces an adaptive mechanism to reduce the overall number of free choice questions.

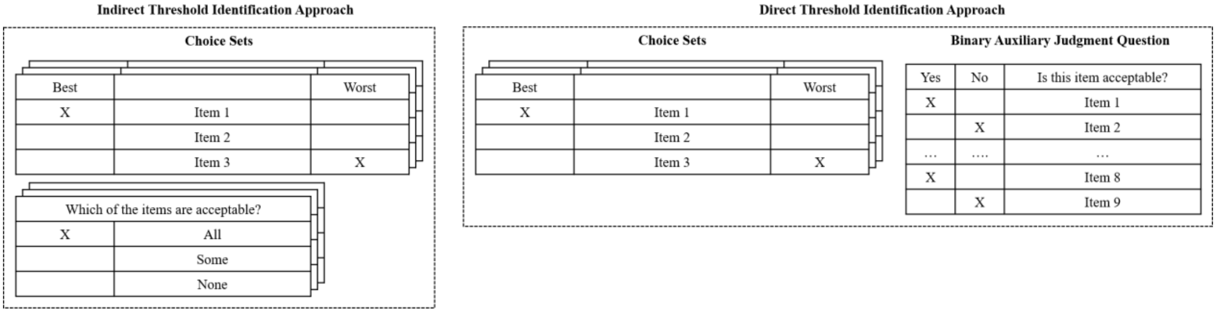
The way in which the no-choice option is integrated into the choice task can influence the frequency with which the no-choice option is chosen (e.g., Wlömert & Eggers, 2016). For instance, Schlereth and Skiera (2017) summarize studies in which the no-choice option was chosen twice as frequently in a dual response setting compared to traditional choice-based conjoint. As a consequence, willingness to pay estimates substantially differed depending on how the no-choice option was included. It is therefore relevant to understand how the introduction of a threshold identification approach affects the results.

2.2 Threshold identification in best-worst scaling

The threshold identification in BWS has so far received little attention. A prerequisite for threshold identification in BWS is the possibility to apply a second criterion to the item ranking. Considering the introductory example, the threshold question which car brands fall into a consumer's consideration set provides complementary information in addition to the obtained relative preference ranking of car brands.

We can distinguish two types of approaches to elicit the additional information for threshold identification in BWS, namely indirect and direct approaches. Indirect approaches share similarities to dual response and ask respondents immediately after each choice task whether the presented items are acceptable or not. For example, Louviere, Flynn, and Marley (2015) ask respondents whether "all", "some", or "none" of the items in a choice set are acceptable. Direct approaches, in contrast, present the respondent with the complete set of items after all best-worst tasks have been completed (Dyachenko et al., 2014; Lattery, 2010). The respondent then has to decide, for each item, whether it is acceptable or not, for instance in a binary auxiliary judgment question (see Figure 1).

Notably, there is no agreement in the literature on which threshold identification approaches to use and how to embed the threshold information in an appropriate model. It is also unknown whether the choice for either of the two methods affects the outcome. We aim to contribute to existing literature on BWS by testing and comparing different threshold identification approaches and developing corresponding models for the multinomial Hierarchical Bayes (HB) estimation.


Indirect Threshold Identification Approach

Choice Sets

Best		Worst
X	Item 1	
	Item 2	
	Item 3	X

Which of the items are acceptable?	
X	All
	Some
	None

Direct Threshold Identification Approach

Choice Sets

Best		Worst
X	Item 1	
	Item 2	
	Item 3	X

Binary Auxiliary Judgment Question

Yes	No	Is this item acceptable?
X		Item 1
	X	Item 2
…	….	…
X		Item 8
	X	Item 9

Figure 1. Illustration of Indirect and Direct Threshold Identification Approaches

3 Methodology

3.1 Problem of identifying a threshold

Before we present the model for HB estimation, we illustrate the underlying threshold estimation 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 (Kuhfeld, Tobias, & Garratt, 1994; Louviere et al., 2015, pp. 16-20), the Count method simply adds a score of 1 for every time an item has been picked as best, and subtracts a score of 1 for every time an item has been picked as worst (Louviere et al., 2015). Consequently, the best-worst scores range between [-r; +r], where r represents the number of repetitions per item. Some of the best-worst scores might equal zero. However, this zero is not suitable to distinguish whether a given item is acceptable or not. Instead, we use the decisions from threshold identification questions to shift the best-worst scale such that a score of zero on the shifted scale coincides with the threshold. Figure 2 illustrates this idea for the case of four repetitions per item.

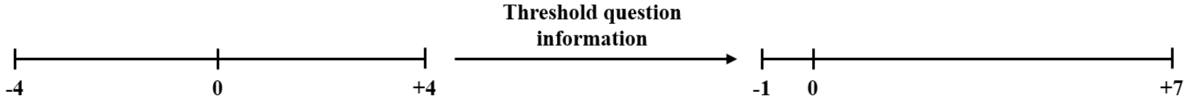


Figure 2. Illustration of Scale Shift

3.2 Threshold measurement

Assume a best-worst study with multiple items in which a subset of three items is 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 acceptable (V1).

We note that 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 “acceptable” or “unacceptable” items. In this paper, we therefore test two modifications: V2 asks explicitly for the number of items in a choice set that are considered acceptable. Respondents need to make an evaluation of all items and thus we obtain information for the middle item. V3 also accomplishes that goal but asks respondents to make a judgment only on the middle item. Last, we implement the binary auxiliary judgment question as a direct threshold identification approach (V4). Table 1 summarizes the versions in terms of their response options.

V1 (Louviere)	*V2 (new)*	*V3 (new)*	*V4 (Lattery)*
For each choice set:	*For each choice set:*	*For each choice set:*	*For each item:*
All items are acceptable	All 3 items are acceptable (3-0)	Middle item is acceptable	Is this item acceptable?
Some items are acceptable	2 items are acceptable, 1 item is not (2-1)	Middle item is unacceptable	□ Yes
None of the items are acceptable	1 item is acceptable, 2 items are not (1-2)		□ No
	All 3 items are unacceptable (0-3)		

Table 1. Three Versions of Indirect Threshold Questions

3.3 Estimation

We describe the mathematical underpinnings of best-worst score and threshold estimation exemplary for V3, and focus on the hierarchical Bayes’ likelihood function. This function consists of three components that respectively pertain to the best, worst, and threshold choices. Let $d_{hij,best}$ [$d_{hij,worst}$] be an indicator whether respondent h chose the jth item in choice set i as best [worst] and v_{hij} be the item’s best-worst score. Further, let $v_{hi,m}$ be the best-worst score of the middle item, $d_{hi,m,thr}$ be an indicator whether the middle item was judged as acceptable, and $v_{h,thr}$ be the threshold. Then the respondent-specific likelihood is:

$$L_h = \prod_{i=1}^{I}\prod_{j=1}^{J}(\frac{e^{v_{hij}}}{\sum_{j=1}^{J} e^{v_{hij}}})^{d_{hij,best}} * (1-\frac{e^{v_{hij}}}{\sum_{j=1}^{J} e^{v_{hij}}})^{1-d_{hij,best}} * \prod_{i=1}^{I}\prod_{j=1}^{J\backslash\{best\}}(\frac{e^{-v_{hij}}}{\sum_{j=1}^{J\backslash\{best\}} e^{-v_{hij}}})^{d_{hij,worst}} * (1-\frac{e^{-v_{hij}}}{\sum_{j=1}^{J\backslash\{best\}} e^{-v_{hij}}})^{1-d_{hij,worst}} *$$

$$\prod_{i=1}^{I}(\frac{e^{v_{hi,m}+v_{h,thr}}}{1+e^{v_{hi,m}+v_{h,thr}}})^{d_{hi,m,thr}} * (\frac{1}{1+e^{v_{hi,m}+v_{h,thr}}})^{1-d_{hi,m,thr}} \quad (h \epsilon H) \qquad (1).$$

The choice probabilities for the worst item are determined based on the set of items that remain after the best item was chosen. Relying on this likelihood specification, the best-worst score of each item as well as the threshold are estimated on a respondent level based on HB.

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 by adding the threshold to the respective best-worst score. Items with a positive [negative] shifted best-worst score can now be categorized as “acceptable” [“unacceptable”], respectively.

The likelihood functions for V1 and V2 are comparable, with the difference that not only the middle item, but also the best and worst item in each choice set must be considered. In case of V4, the likelihood accounts for the fact that threshold decisions are made on an item instead of choice set level.

3.4 Simulation

We conducted a simulation study in which we sampled synthetic choices from normally distributed preferences with standard deviation σ. The results show that the version which asks for the most information recovers simulated parameters with the smallest error under all conditions. While the simulation supports the intuitive insight that asking for more information, as in V2, outperforms asking for less information, it is yet unknown how the threshold questions will perform in an empirical setting and how repondents react to the threshold questions.

		RMSE			*MAE*		
		V1	*V2*	*V3*	*V1*	*V2*	*V3*
σ	*High*	8.54	7.97	9.07	6.37	5.99	7.33
	Low	6.07	5.60	7.30	4.78	4.30	5.81

Table 2. Error Rates in Recovery of Simulated Threshold Parameters (Rescaled to [0; 100])

4 Empirical Studies

4.1 Study design

We conducted two studies to examine whether the type of threshold question included in BWS has an impact on respondents’ choice behavior and resulting parameter estimates. In each study, we evaluated nine items and employed a balanced incomplete block design (BIBD) with twelve choice sets of three items each such that every item appeared four times across choice sets (Louviere et al., 2015; Louviere et al., 2013). Subjects were randomly allocated to one of four survey versions, which only differed in their threshold question. We implemented the threshold questions of V1 to V4.

In Study 1, respondents were asked to express their opinion on non-profit organizations, presented alongside their logo and mission statement. The organizations were selected from a broad spectrum of fields such that they would expectedly differ in their broadness of appeal.

In each choice set, respondents were asked to select the organization they would “most likely” and “least likely” donate to. In the threshold question, respondents had to decide whether they are “rather likely” or “rather unlikely” to donate to the respective organizations. Respondents were instructed to choose as if they were donating themselves. We obtained a sample size of 496 [V1: 129, V2: 104, V3: 130, V4: 133]. To reduce hypothetical bias, we introduced an incentive-compatible design in which actual donations were distributed in accordance with respondents’ indicated preferences (Dong, Ding, & Huber, 2010). We randomly selected one out of every ten respondents and committed to distributing a donation amount of €10 among each winner’s most preferred organizations. Organizations below the identified threshold did not receive any donations.

In Study 2, we conducted a replication including V1 to V3 to see whether our results from Study 1 still hold in a non-incentive-compatible setting. Specifically, respondents were asked to express their opinion on adopting nine different environment-protecting actions. We obtained a sample size of 306 [V1: 97, V2: 104, V3: 105].

4.2 Comparison of survey versions

We examine whether different ways of including the threshold question have a significant effect on response behavior. In a first step, we investigate whether the type of threshold question leads respondents to behave more or less consistently in making best and worst choices. We determined a consistency score for each respondent as the sum of squared best-worst scores of all included items (Louviere et al., 2015, p. 29), which ranges between 0 and 60. Conducting an ANOVA, differences among consistency scores across versions are found to be insignificant in either study ($F_1(3, 492) = 2.23$, $p_1 = .083$; $F_2(2, 303) = 0.15$, $p_2 = .863$).

Furthermore, we test whether respondents’ consistency in making threshold decisions differs across types of threshold questions. Respondents make an inconsistent threshold decision if they declare in one choice set that they would “rather likely” donate to a certain organization or engage in a certain action and in another choice set that they would “rather unlikely” do the same. On a respondent level, consistency is measured as the percentage of items that, out of the presented nine, were consistently classified. We note that there are significant differences in respondents’ threshold consistency among V1 to V3 in both studies ($F_1(2, 360) = 20.24$, $p_1 < .001$; $F_2(2, 303) = 23.60$, $p_2 < .001$). A t-test with Bonferroni correction reveals for both studies that V2 yields a significantly lower threshold consistency than V1 ($p_{1,2} < .001$) and V3 ($p_{1,2} < .001$), respectively. The survey versions also differ significantly in the percentage of respondents that were consistent in their threshold decisions across all items (X_1^2(2, N = 363) = 23.42, p_1 < .001; X_2^2(2, N = 306) = 29.86, p_2 < .001). V3 performs significantly better than V1 (p_1 = .044; p_2 < .001) and V2 ($p_{1,2}$ < .001) in both studies. Furthermore, V1 performs significantly better than V2 (p_1 = .020) in Study 1.

A final question is whether there are actual and perceived differences in the level of effort respondents need to expend to complete the different survey versions. As a first proxy, we measured the time respondents took to make best and worst choices and answer the threshold questions. An ANOVA reveals no significant differences in response time among survey versions (F_1(3, 492) = 1.73, p_1 = .161; F_2(2, 303) = 0.60, p_2 = .551). In Study 1, respondents also provided their own estimate of completion time. There are no significant differences across survey versions with regard to this measure (F_1(3, 489) = 0.12, p_1 = .948). Cognitive ease of completing the different survey versions was measured using a four-item semantic differential scale (adapted from Bettmann, John, and Scott (1986, p. 319); α = .77). An ANOVA reveals that perceived complexity is the same across versions in both studies (F_1(3, 492) = 1.56, p_1 = .199; F_2(2, 303) = 0.88, p_2 = .416).

	Study 1				***Study 2***		
	V1	*V2*	*V3*	*V4*	*V1*	*V2*	*V3*
Consistency Index: $\frac{1}{H}\sum_{h=1}^{H}\sum_{i=1}^{9} BW_{hi}^2$	57.44	57.87	58.23	58.66	56.74	56.37	56.30
Consistent Threshold Choices	89.23%	81.73%	92.22%	-	84.65%	73.72%	89.42%
Perfectly Consistent Respondents (Threshold)	41.09%	24.04%	55.38%	-	26.80%	18.27%	52.38%
Median Completion Time BWS [s]	238	244	228	223	301	321	319
Mean Perceived Completion Time BWS [min]	7.19	7.06	6.97	7.08	-	-	-
Mean Cognitive Ease Score[a]	5.38	5.55	5.50	5.25	5.06	5.26	5.04

[a] Composite of items *simple/complex, easy/hard, easy to follow/hard to follow, not difficult to complete/difficult to complete*; items measured on a scale from 1 to 7; one item was reverse coded

Table 3. Comparison of Best-Worst Survey Versions

4.3 Estimation results

We determined the best-worst scores for each of the nine items as well as the threshold on the individual respondent level using HB, and subsequently averaged across respondents.

Study 1	*V1*	*V2*	*V3*	*V4*	***Study 2***	*V1*	*V2*	*V3*
SOS Kinderdorf	2.96	2.72	2.72	2.87	Reusable shopping bag	1.86	1.79	1.72
WWF	2.61	2.61	2.61	2.37	Hang laundry to dry	0.98	1.00	1.37
PETA	0.52	0.08	0.10	0.40	Buy regional products	0.30	-0.04	-0.43
Robin Hood	-0.17	0.32	-0.22	0.13	Reusable coffee mug	0.01	0.03	0.10
Grüne	-0.44	-0.75	-0.56	-0.80	Bicycle/public transport	-0.18	0.02	0.16
Bundeswehr	-0.88	-0.87	-0.92	-0.73	Turn off devices	-0.43	-0.48	-0.41
CDU	-0.90	-0.63	-0.70	-0.73	Buy glass bottles	-0.50	-0.77	-1.26
Kirche in Not	-1.63	-1.38	-0.94	-1.67	Renewable energy	-0.95	-0.97	-0.23
Extinction Rebellion	-2.06	-2.11	-2.10	-1.84	Less meat	-1.09	-0.58	-1.02
Threshold	-1.29	-1.15	-0.83	-1.15	Threshold	1.67	1.70	2.22

Table 4. Mean Parameter Estimates (Effect-coded Best-Worst Scores and Threshold)

We plotted the relation of mean *shifted* best-worst scores between survey version pairs (Figure 3). For Study 1, the mean shifted best-worst scores scatter around a line through origin with a slope of one in all survey version pairs. This is an indication that the nature of the threshold question does not generally influence respondents' choices. Further, as all items are situated in either the first or third quadrant in each graph, there is agreement among survey versions as to the organizations which the average subject would (two organizations) or would not donate to (seven organizations). Similar observations apply to Study 2. Here, all versions agree that the average respondent is rather likely to adopt all nine environment-protecting actions.

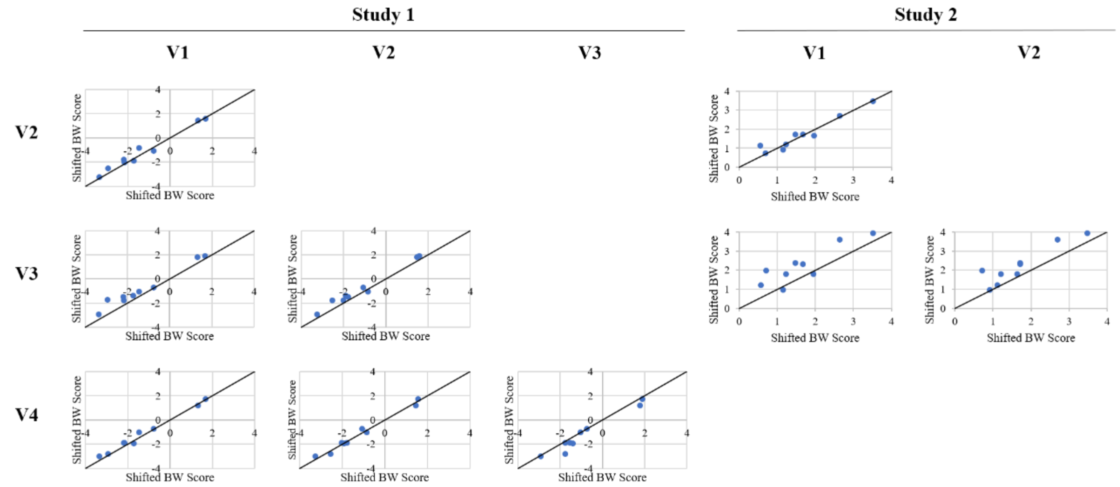


Figure 3. Relation between Mean Shifted Best-Worst Scores from Different Survey Versions

Finally, we translated the individual preferences from Study 1 into donation amounts for each organization. The resulting donation distribution, normed to 100 respondents, is presented in Table 5. We conducted an ANOVA for each of the nine organizations and found no significant differences in donation amounts among survey versions for any organization. However, V3 and V4 yield comparably higher average rank correlations.

Donation amount, normed to 100 respondents, each donating €10	*V1*	*V2*	*V3*	*V4*
SOS Kinderdorf	329.18	314.20	288.20	343.43
WWF	283.83	313.83	311.87	278.09
PETA	89.69	74.31	99.60	105.51
Grüne	51.22	27.88	36.95	41.57
Robin Hood	49.10	79.93	60.26	76.69
Bundeswehr	38.17	44.44	36.50	34.01
CDU	26.78	52.07	45.13	50.88
Kirche in Not	26.23	29.07	31.77	21.05
Extinction Rebellion	12.78	16.21	12.79	18.69

Note. Average rank correlations vis-à-vis the remaining three versions: V1: .87, V2: .88, V3: .93, V4: .94

Table 5. Donation Amounts by Organization, Normed to 100 Respondents

5 Conclusion

The biggest surprise is likely that, in contrast to traditional DCEs, in which the way that threshold idenfication questions are included can substantially affect the location of the measured threshold, for BWS, the results are robust. This robustness is observed in two empirical studies, one with and one without incentive alignment. Yet, we see differences between the approaches, in particular with respect to respondents' consistency in making threshold decisions. We therefore advocate for the approach that asks respondents to provide the least information and yields the highest threshold choice consistency, namely V3.

We encourage future research to build on and extend our findings by investigating ways to benchmark the threshold against an external reference value and adapting the HB model to account for effects of choice order and certainty.

6 Literature

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

Brazell, J. D., Diener, C. G., Karniouchina, E., Moore, W. L., Séverin, V., & Uldry, P.-F. (2006). The no-choice option and dual response choice designs. *Marketing Letters, 17*(4), 255-268.

Brynjolfsson, E., Collis, A., & Eggers, F. (2019). Using massive online choice experiments to measure changes in well-being. *Proceedings of the National Academy of Sciences, 116*(15), 7250-7255.

Dhar, R. (1997). Consumer preference for a no-choice option. *Journal of Consumer Research, 24*(2), 215-231.

Dhar, R., & Simonson, I. (2003). The effect of forced choice on choice. *Journal of Marketing Research, 40*(2), 146-160.

Dong, S., Ding, M., & Huber, J. (2010). A simple mechanism to incentive-align conjoint experiments. *International Journal of Research in Marketing, 27*(1), 25-32.

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

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.

Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research, 31*(4), 545-557.

Lattery, K. (2010). Anchoring maximum difference scaling against a threshold - Dual response and direct binary responses. *Proceedings of the 15th Sawtooth Software Conference,* 91-106.

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.

Schlereth, C., & Skiera, B. (2017). Two new features in discrete choice experiments to improve willingness to pay estimation that result in new methods: Separated (adaptive) dual response. *Management Science, 63*(3), 829-842.

Wlömert, N., & Eggers, F. (2016). Predicting new service adoption with conjoint analysis: External validity of BDM-based incentive-aligned and dual-response choice designs. *Marketing Letters, 27*(1), 195-210.