Heterogeneity in the Consistency of Best-Worst Scale Responses

Luke Greenacre[1], Steven Dunn, and Ana Mocanu

Ehrenberg-Bass Institute and School of Marketing

University of South Australia, SA 5000, Australia

[1] Corresponding author:

luke.greenacre@unisa.edu.au, Fax: +61 8 8302 0442

University of South Australia, Adelaide, SA 5000, Australia

Abstract

Best-Worst Scaling is one of the dominant measurement approaches in choice experimentation. When employed it provides substantial information on peoples' preferences without making choice tasks prohibitively long. Although, one concern with this method is that peoples' selection of a best may not reflect the same preferences as when a worst is selected. Research into such an inconsistency between best and worst responses has found it to be a non-trivial and persistent problem. This research further investigates these inconsistencies and finds that they can largely be attributed to a relatively small group of people in the sample who do not anchor their worst responses onto their best responses as literature suggests they would. In fact, 25% of the participants in a sample account for between 50 and 60% of the inconsistent responses recorded. The presence of this group, and their disproportionate contribution to the number of inconsistencies in best and worst responses provides strong evidence that there is heterogeneity in how consistently people formulate responses in best-worst tasks. Recommendations are made regarding how to accommodate this phenomenon in utility based choice models so that better predictions of choices can be made.

Highlights

- This paper finds that most of the inconsistency between best and worst responses in Best/Worst experiments can be attributed to a sub-set of the sample measured
- Approximately 25% of a research sample accounts for 50-60% of the inconsistencies between best and worst responses
- Using two contexts, we show there is no clear reason as to why these people are inconsistent. They are attentive to the task and take the same amount of time to complete the experiment.
- Analysis needs to accommodate this heterogeneity in Best/Worst choice consistency for accurate prediction using Best/Worst data. Recommendations for analysis are made.

Keywords

Choice, experiment, best-worst, scale, heterogeneity, consistency

1. Introduction

The interest in Best-Worst (BW) scaling has increased dramatically over the last few years (Flynn, Louviere, Peters, & Coast, 2007; Louviere et al., 2013; Marley & Louviere, 2005). BW scaling overcomes many of the limitations of other measurement methods used in survey research, such as ranking and rating scales. As an extension of the standard discrete choice experiment, a BW experiment asks individuals to choose both their top and bottom ranked alternatives in a choice set. The alternatives used can be things such as political parties (Remaud & Gillan, 2007), policies and opinions (Finn & Louviere, 1992; Jones et al., 2013), means of transport (Outwater et al., 2013), medical treatments (Flynn, Louviere, Peters, & Coast, 2007), consumer products (Cohen, E., 2009; Louviere et al., 2013), or any other object that people may naturally choose amongst. The Best-Worst format though provides substantially more preference information than a standard choice experiment.

With the introduction of a new measurement method, natural concerns arise about potential new and unknown biases or errors that the method may introduce into data sets. Previous literature has expressed concern that the best and worst responses elicited from people in a BW experiment may not reflect the same underlying preferences, or have differing scales (Flynn, Louviere, Peters, & Coast, 2010; Louviere & Eagle, 2006). This lack of consistency between best and worst responses would make analysis of this data more difficult. Research has identified that inconsistencies between best and worst responses are generally small but certainly non-trivial (Mueller Loose & Lockshin, 2013).

In a parallel literature, preference and response *heterogeneity* have been raised as a potential source of error in analysis of BW data, and in choice experiments in general (Cardell, 1997; Flynn et al., 2010; Hutchinson, Kamakura, & Lynch, 2000). Whether there are differences between individuals in preferences and in how they make choices has important implications for analysis and interpretation.

What has not been considered is that heterogeneity and BW consistency may in fact be related. This paper argues that for a sub-group of people, there is a greater difference between how they formulate their best responses and their worst responses. We explore whether there is heterogeneity across individuals in how consistently they select best versus worst alternatives. What we find is that this is indeed the case, a non-trivial sub-group of participants make choices of best that are not consistent with their choices of worst.

2. Literature

2.1. Best-Worst Scaling

BW scaling is a generalization from discrete choice experimentation. In a standard choice experiment the participant is asked to select the alternative in each choice set that they prefer the most. BW scaling forces people to choose the most preferred (best) and the least preferred (worst) alternatives from a choice set. Through this elicitation across numerous choice sets, substantial information about peoples' preferences is gathered (Louviere et al., 2013).

The BW approach is free of the scale biases present in popular rating and ranking approaches (Louviere & Islam, 2008). BW scaling produces a ratio level scale that allows for improved comparisons across demographic segments (Cohen, 2009). When individuals evaluate a set of objects, their extreme choices of best and worst alternatives are expected to be more reliable than choices among middle items. This provides an improvement in response reliability over the ranking of all alternatives. It also comes more naturally to people to select what they like most and least of a subset of alternatives, than to rate their preference for alternatives from 0 to 10 for example. It has been argued that rating scales have little, if any, equivalence in the typical day-to-day decision-making process (Louviere et al., 2013).

Extant literature recognizes three cases of BW measurements. The first case is the object case. In this case, individuals are asked to choose the best and worst alternative in a choice set (Marley & Louviere, 2005). Each alternative is a simple object that is expected to be holistically evaluated. For example this could simply be the choice among named brands. The second and third cases are extensions of the first. In case two, sometimes called the profile case, individuals choose from alternatives that have different profiles described as combinations of attributes based on an underlying design. These profiles are presented one at a time and the best and worst *attribute levels* within each profile are chosen (Flynn, Louviere, Peters, & Coast, 2007). For example, the each brand along with its ingredients, if a food, would now be presented individually, with participants selecting the best and worst ingredients for each brand. The objective of case two is to identify the critical attributes or features driving peoples' choices. In case three, individuals choose the best and worst designed *alternatives* from various choice sets based on an underlying design (Marley & Pihlens, 2012). That is, they choose an object from those shown, much like in case one. The difference to case one is that the alternatives are designed as an experimental combination of attributes and levels. For example, the choice would now be among a number of branded food products, with those products being both named and having their specific ingredients listed. Case three is widely used, the most elaborate, and most powerful in an applied setting. Case three allows for testing of whole objects and the formulation of predictions of population level outcomes. In this paper, we focus on case three.

2.2. Consistency and Heterogeneity in Best and Worst Responses

Since the early development of BW scaling a number of theoretical approaches have been considered for reconciling the best and worst responses into a single measure of people's preferences (Marley, Flynn, & Louviere, 2008; Marley & Louviere, 2005; Marley & Pihlens, 2012). Present in all of these approaches is an acknowledged concern that there may be a lack of consistency between best and worst responses in a BW experiment. Two of the greater concerns are that people may formulate different preferences when prompted for a best alternative than for a worst alternative, and/or that their responses for best may be on different utility scales to that of worst responses. Findings from the decision framing literature further compounds these concerns, as it has been discovered that framing decisions as selections versus rejections elicits different preferences (Laran & Wilcox, 2011; Shafir, 1993). The potential parallels between selections and rejections, and best and worst responses are obvious, hence concerns that such findings may extend into BW.

Testing has shown that there tends to be agreement between people's best and worst responses (Mueller Loose & Lockshin, 2013). Although few instances of extreme discrepancy have been found small and persistent discrepancies are identified in almost all applications of BW scaling. While small, they are certainly non-trivial as they can have considerable impact on estimation of utility based models (Marley & Pihlens, 2012). Even small discrepancies can lead to inaccurate predictions of population level outcomes.

The *source* of these discrepancies has largely been ignored in the literature. Most applications of BW experimentation are across large samples of participants. By aggregating the BW scores of people we are generally assuming the inconsistencies in the best and worst responses are a feature of the sample as a whole. What we argue though is that the discrepancies between best and worst responses seen across the whole sample can largely be attributed to a relatively small group of people that are less able to generate consistent BW responses.

Heterogeneity in the sample in the ability to formulate consistent best and worst responses would suggest another feature of decision making that would need to be included in analysis employing BW data. Heterogeneity has been raised as a potential source of the inconsistency between selections and rejection in the framing literature (Hutchinson et al., 2000). Such findings lead us to question whether concerns about heterogeneity are also warranted here.

Some people may be better at formulating best responses that are consistent with their worst responses. It is generally thought, although not explicitly stated in the literature, that worst responses are anchored to best responses, which are usually prompted for first. Such thinking arises as BW is an extension of a standard choice experiment, where a selection (choice of ‘best’) is the primary response type. Anchoring implies that a single decision making process is activated, with the prompting for the worst alternative being a mirror of the best response. The results in literature largely support this occurring. Best and worst responses are consistent for the most part (Marley & Pihlens, 2012). Some people may have weaker anchoring to the best, leading to the prompt for worst responses to activate a different decision making process to that of best. Weak anchoring would thus introduce a subtle framing effect into BW data for those people.

The presence of a group of people that are more subject to a framing effect between the best and worst responses could considerably degrade the usefulness of BW data for explaining and predicting their behaviour. Having even a small group of respondents being less consistent in their responses could introduce substantial error into a choice model of a larger population. Even a small group of people could be introducing large numbers of inconsistent choices into the data set. This account of heterogeneity in best worst response consistency thus needs to be tested.

3. Methodology

3.1. Experiment design

A sequence of three choice experiments is used in this study, each forming one experimental condition. In the first experimental condition, participants indicate products they are most likely to purchase (the 'best') from pairs of products (see Figure A.1 for an example). The second experimental condition replicates the first except it asks participants to select the products they are least likely to purchase (the 'worst') (see Figure A.2 for an example). In comparing the responses from these two conditions, we can identify people with greater (in)consistency between their best and worst responses.

[Figures A.1 and A.3 about here]

A third experimental condition in this study uses the same products as in the first two, but they are now presented in a standard BW experimental format. In this format, participants identify the most and least likely product they would purchase from sets of four products at a time (see Figure A.3 for an example). By including the first two experimental conditions, we can effectively identify the more and less consistent individuals, specifying the source of heterogeneity in the consistency between best and worst responses. We can then evaluate whether this heterogeneity persists into the more standard BW experiment format using the final experimental condition. Finding a less consistent group and confirming that their responses in the BW experiment condition are different to that of the population would show that BW response heterogeneity is likely present in data collected using BW scaling.

The first two experimental conditions that only show pairs of products are needed because such best-worst heterogeneity is exceptionally difficult to detect in a standard BW experimental approach (Marley & Louviere, 2005). Standard BW experiments include a minimum of three alternatives in a choice set, typically including between 4 to 8 alternatives in a choice set. The experiment then only collects data points on two of those alternatives and we do not know the preference orderings for the remaining alternatives. The presence of such missing information means that we cannot define the inverse of the best, as being empirically equivalent to the worst (Marley & Louviere, 2005). This lack of equivalence makes accurate assessment of the consistency between best and worst responses difficult. In a binary choice experiment however, the inverse of the best should *perfectly* match the worst, providing us a much more accurate method to assess the consistency of the responses provided by participants.

3.2. Product categories and attributes

Two product categories were tested in this research, juice and pizza restaurants. These two categories were selected based on their successful use in BW and choice-based experiments (Louviere & Islam, 2008; Louviere, Islam, Wasi, Street, & Burgess, 2008). They are also relatively simple choice objects that can be easily evaluated by the general population. The juice and pizza restaurant alternatives were designed using different Orthogonal Main Effects Plans that resulted in 16 products being available for selection in each product category. Seven attributes were used to characterise each juice: flavour (orange, apple), servings per pack (3, 6, 9, 12), price per serve ($0.30, 0.40, 0.50, 0.60), percentage of real juice content

(40%, 70, 100), whether it was concentrate (or not), added vitamins (vitamin C, none added), and sugar content (sweetened or not). The pizza restaurant alternatives used: price ($10, 12, 14, 16), delivery time (20 min, 30, 45), delivery guarantee (free if late, none), toppings (3-4, 4-6), garlic bread (free, $2), wings (2 free then $1, $1), salad (free, $2), and firing (wood, oven).

The first two binary choice experiment conditions that have people evaluating pairs of alternatives only were constructed using a combinatorial design. The 16 products were organised into all pair-wise combinations producing 120 choice sets that the participants would consider. The full 120 pair-wise combinations was used to obtain as full preference information as possible to ensure that any subsequent inconsistencies in conditions could be attributed to preference and not to the design. The first condition had participants choosing the 'best' alternative for each of the 120 pairs. The second condition had participants then choose the 'worst' alternative for each of the 120 pairs. The third BW experimental condition that had four alternatives in each choice set was constructed using a Balanced Incomplete Block Design (BIBD). This is a typical design for a BW experiment. The BIBD produced 20 choice sets with four alternatives per choice set. To control for potential confounding effects the order the three conditions were blocked resulting in six possible survey layouts. Participants were randomly assigned to a layout.

3.3. Analysis

The sample for the juice category version of the experiment consisted of 171 residents of the US (70 males, 37% aged 25-34 years). The sample for the pizza restaurant category was 268 US residents (122 males, 40% aged 25-34 years). Both samples were sourced from the online panel provider Mechanical Turk, a service of Amazon.com. Incentives were provided for participation to the approximate value of US$2. Attention tasks were included in the survey to confirm participants were correctly attending to the survey. These attention tasks involved five simple mathematics problems that the participant had to get correct for us to include their data. Whether correct or not they were able to complete the survey and receive payment, ensuring they were not incentivised to excessively attend to this task. All participants satisfactorily completed these tasks in the juice category, and two people were excluded prior to any reporting based on these tasks in the pizza restaurant category.

4. Results

Before looking for any heterogeneity in response consistency it is important to determine whether there is any inconsistency in best-worst responses as per standard practice with BW scaling. One of the most rigorous tests of consistency between the best and worst responses in a BW experiment is to regress the average choice frequency for the best responses on the square root of the ratio of average best to average worst responses for each choice alternative (Lee, Soutar, & Louviere, 2008; Mueller Loose & Lockshin, 2013). Regression coefficients approaching 1 would indicate consistency in responses.

[Figures B.1 and B.2 about here]

Figure B.1 and B.2 show markedly different outcomes from this test. Both regressions have a good r-squared ($R^2_j = .96$, $R^2_p = .92$), and both categories exhibit acceptable levels of consistency between the best and worst responses in the BW experiment (null: $\beta = 1$; $\beta_j = 0.87$, $t = -2.63$; $\beta_p = 1.13$, $t = 1.45$). In both cases there is some level of inconsistency between the best and worst responses with the juice category in particular seeing a significant deviation from a coefficient of one. Having one category with significant deviations and one without gives us the opportunity to see if the proposed heterogeneity is present even in seemingly acceptable data sets.

To detect whether this lack of consistency could be attributed to heterogeneity in response consistency we compare the responses from the two binary choice experiment conditions. For the purpose of all analysis, the responses in the first two choice experiment conditions among pairs of alternatives the *worst responses have been reverse coded* to allow for easy comparison to the binary best responses. This reverse coding involved taking the alternative *not chosen* from the pair and using that as the basis for analysis.

[Figures C.1 and C.2 about here]

Figures C.1 and C.2 show the average response frequency for each of the alternatives in the two binary choice experiment conditions. We can see that the best and worst (reversed) responses reflect very similar preference structures. Indeed the best and worst responses are highly correlated for both (r_j = .998, p < .01; r_p = .999, p < .01). We can also observe that the response frequencies do not match perfectly. In fact, the worst responses appear to generally underestimate the preference for higher preference objects and overestimate the preference for lower preference objects relative to best responses. To illustrate, take the most extreme columns in Figure C.1, the least preferred juice alternative has a ‘best’ frequency of 4.63, less than the (reversed) ‘worst’ frequency of 5.11; and the most preferred alternative has a ‘best’ frequency of 12.12, more than the (reversed) ‘worst’ frequency of 11.73. The same pattern is present for the surrounding columns in the Figure, and in Figure C.2. Some people, at least some of the time, did not formulate best responses that were perfectly consistent with their worst responses.

We can compare each choice set from the two binary choice conditions for each individual and determine the number of times that they made an inconsistent response. The responses for each choice set from the binary choice condition where the person indicated the ‘best’ were directly compared to the responses from the same choice set where the person indicated the ‘worst’. Where the ‘best’ choice is the precise opposite of the ‘worst’, that is where the complement of the best is the worst, this was counted as a consistent behaviour. Where the two choices did not match it was coded as an inconsistency. Using this comparison we can measure each participant’s inconsistency on a scale from 0 up to 120, the number of choice sets in the binary choice experiment conditions. The distribution of these inconsistencies is shown in Figures D.1 and D.2.

[Figures D.1 and D.2 about here]

The distributions show heavy right skew, indicating a relatively small number of people have a large number of inconsistent responses. The top twenty five percent of the sample in the juice and pizza restaurant categories provide inconsistent responses in *at least* 33 and 28 of the 120 choice sets respectively. This smaller group (25% of the sample) account for the disproportionate quantity of 59% and 52% of the total number of inconsistent responses found in the juice and pizza categories. Based on these findings, we define two groups of people for both samples - a more consistent group (the bottom 75% of the sample) and a less consistent group (the top 25% of the sample).

To confirm that the less consistent group did not emerge due to task inattention we compared the completion times for the experiment between the two groups found. If inattentive we would expect the less consistent group to have completed the experiment faster. In this case though, no significant difference in completion time was found for either the juice category (*F* = 8.964, *t* = .137, *p* = .891) or the pizza category (*F* = 5.509, *t* = .503, *p* = .616).

The results thus far, demonstrate that there is a subset of people that produce less consistent responses for both best tasks and worst tasks when used separately in our binary choice experiment conditions. A standard BW experiment, equivalent to our third condition, that collects the best and worst responses at the same time has a clearly different structure though as there is a greater opportunity for worst responses to be anchored to the best response. Consequently, we need to examine whether the heterogeneity in response consistency persists into the BW data as well.

Taking the data from the BW experimental condition, we split the sample into the previously identified higher and lower consistency groups. For each group we test the response and scale consistency by again regressing the average choice frequency for the best responses on the square root of the ratio of average best to average worst responses for each choice alternative (Lee et al., 2008; Mueller Loose & Lockshin, 2013).

[Figures E.1 and E.2 about here]

Figure E.1 for the Juice category shows that the less consistent group retains its inconsistency in the best-worst data. The regressions have good r-squares for both groups (R^2_{jless} = .94, R^2_{jmore} = .94), only the more consistent group retains the reasonable levels of consistency between the best and worst responses seen at the aggregate level (null: β = 1; β_{jless} = 0.67, *t* = -7.11; β_{jmore} = 0.93, *t* = -1.18). The pizza restaurant category in Figure E.2 sees similarly good r-squared results (R^2_{pless} = .94, R^2_{pmore} = .84) and it is clear that the less consistent group (null: β = 1; β_{pless} = .90, t = -1.60) was masking issues present with the more consistent group in this case (null: β = 1; β_{pmore} = 1.39, t = 2.39). For this product category, the more consistent group was not necessarily better, just markedly different from the less consistent group.

From the results, we can see that the group that had greater consistency in the binary experiment conditions also tends to have much more consistent responses, or at least vastly different responses to the population, in the BW experimental condition. The reverse is also true for the lower consistency group. This result provides evidence that there is heterogeneity in how consistently people formulate their best and worst responses. In the present data sets, the less consistent group were clearly attending to the task. The adequate completion of the attention tasks included into the method confirms that. Fatigue is also unlikely to be a factor because the less consistent group were not over-represented in one particular ordering of the experiments designed into the survey. The less consistent group in the juice category was exposed to the standard BW experiment first in the sequence approximately 38% of the time, and last 29% of the time. In the pizza restaurant category, these are 30% and 34%, respectively.

The differences in responses from the first two choice experiments amongst the pairs of alternatives, where framing effects have been demonstrated in detail, persist in the data collected in the BW experiments. Why the less consistent group responds differently from the more consistent group can be difficult to determine, and indeed, there may be *no systematic cause* of this. It can be stated though that the less consistent group does not anchor their worst responses onto their best responses as is implied in the literature.

5. Discussion and Conclusions

The results show that there is heterogeneity in how consistently people formulate best and worst responses. A smaller sub-group of people produce a disproportionately large quantity of inconsistent responses when making best and worst choices as identified in the binary experiments. This sub-group continues this behaviour when undertaking standard BW experiments.

It is generally believed that best and worst responses remain relatively consistent in BW experiment types (Marley et al., 2008; Marley & Louviere, 2005; Marley & Pihlens, 2012). Literature implies that worst responses are anchored onto best responses. Such anchoring would see only a single decision process activated, with the worst response being a mirror of the best response. Little theoretical work has been done in this area though, and evidence from other literatures in decision framing suggest that differences in elicitation methods will activate different decision making processes and different preferences (Laran & Wilcox, 2011; Shafir, 1993). The choice modelling literature has found evidence that best and worst responses are generally consistent in BW experiments (Mueller Loose & Lockshin, 2013). Testing shows that when examined across a large sample the discrepancies between best and worst responses are persistent and not trivial (Marley & Pihlens, 2012).

What this research has found is that a large proportion of the inconsistencies in best and worst responses in BW data can be attributed to a relatively small but substantial group of people. This smaller group of people do not anchor their worst responses onto their best responses as literature suggests. The presence of this smaller group, and their disproportionate contribution to the number of inconsistencies in best and worst responses provides strong evidence that there is heterogeneity in how consistently people formulate such responses.

5.1. Implications

The implications of these findings for analysis are important to consider. The inconsistencies between best and worst responses have largely been viewed as relatively small, albeit not non-trivial, in the literature (Mueller Loose & Lockshin, 2013). This view arose as most testing of best and worst response consistency has been undertaken at the aggregate level, where such inconsistencies are likely to be subsequently attributed to randomly distributed error when modelling. The fact that there is a sub-group in samples that accounts for most of this error indicates a need to parameterise these individual differences in response behaviour.

Individual level modelling (or, latent class modelling) presents the easiest mechanism to resolve this estimation problem (see for example, Huber, 1998). By estimating models for each individual we are able to generate better performing models for the class respondents that do not suffer from differences in how they formulate best and worst responses. While a necessary first step, individual level modelling alone is insufficient. Individual level models will still not perform well for those respondents that inconsistently make best and worst choices.

If this sub-group of people is in fact activating more than one decision-making process due to the framing of choice as bests versus worsts, any individual level models of these individuals will likely perform quite poorly. Thus, the nature of the prompt used to generate each data point needs to be incorporated. Introducing an endogenous variable into models that captures whether a data point was generated via a best or worst response in one obvious way to allow such parameterisation (Dong & Lewbel 2012). How precisely a researcher may wish to formulate that analysis will depend on the assumptions he or she is willing to make about the choice processes and elicitation (Angrist, 1999). What this paper has shown is that it *is* necessary to include such parameterisation of response type into models of choice behaviour when a BW task is employed. Without such inclusion models estimated are likely subject to greater prediction error than is necessary.

5.2. Future research

More research is needed into the relationship between best and worst responses within the choice modelling literature. The psychological processes that are activated when a person undertakes a BW task are poorly understood. Presently it is implied that worst responses are anchored to best responses, as BW scaling is an extension of a standard choice experiment where just the best is chosen. For some people this is clearly not the case. Understanding the nature of the processes that this group of people are activating, and why they are not anchoring their decisions will allow us to more appropriately design BW experiments and model BW data.

References

Angrist, J. D. (1999). Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice. *Journal of Business and Economic Statistics*, *19,* 1-16.

Cardell, N. S. (1997). Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity Variance Components Structures for the Extreme-Value and Logistic Distributions. *Econometric Theory, 13,* 185-213.

Cohen, E. (2009). Applying best-worst scaling to wine marketing. *International Journal of Wine Business Research, 21,* 8-23.

Dong, Y., & Lewbel, A. (2012). A simple estimator for binary choice models with endogenous regressors. unpublished working paper.

Finn, A., & Louviere, J. (1992). Determining the appropriate response to evidence of public concerns: The case of food safety. *Journal of Public Policy & Marketing, 11,* 12-25.

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

Flynn, T. N., Louviere, J., Peters, T. J., & Coast, J. (2010). Using discrete choice experiments to understand preferences for quality of life. Variance-scale heterogeneity matters. *Social Science and Medicine*, *70,* 1957–1965.

Huber, J. (1998). Achieving Individual-Level Predictions from CBC Data: Comparing ICE and Hierarchical Bayes. *Sawtooth Software Research Paper Series*, Sawthooth Software, WA.

Hutchinson, J. W., Kamakura, W. A., & Lynch, J. G., Jr,. (2000). Unobserved heterogeneity as an alternative explanation for ‘reversal’ effects in behavioral research. *Journal of Consumer Research*, *27,* 624-344.

Jones, A.K., Jones, D.L., Edwards-Jones, G., & Cross, P. (2013). Informing decision making in agricultural greenhouse gas mitigation policy: A Best–Worst Scaling survey of expert and farmer opinion in the sheep industry. *Environmental Science & Policy, 29,* 46-56

Laran, J., & Wilcox, K. (2011). Choice, Rejection, and Elaboration on Preference-Inconsistent Alternatives. *Journal of Consumer Research, 38*, 229-241.

Lee, J. A., Soutar, G., & Louviere, J. (2008). The Best–Worst Scaling Approach: An Alternative to Schwartz's Values Survey. *Journal of Personality Assessment*, *90*, 335-347.

Louviere, J., & Eagle, T. (2006). Confound It! That Pesky Little Scale Constant Messes Up Our Convenience Assumptions, *Sawtooth Software Conference*. Florida: Sawtooth Software.

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,* 292-303.

Louviere, J., & 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*, *i* 903-911.

Louviere, J., Islam, T., Wasi, N., Street, D., & Burgess, L. (2008). Designing Discrete Choice Experiments: Do Optimal Designs Come at a Price?. *Journal of Consumer Research*, *35,* 360-375.

Marley, A. A. J., Flynn, T. N., & Louviere, J. (2008). Probabilistic models of set-dependent and attribute-level best–worst choice. *Journal of Mathematical Psychology*, *52*, 281–296.

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

Marley, A. A. J., & Pihlens, D. (2012). Models of best-worst choice and ranking among multi-attribute options (profiles). *Journal of Mathematical Psychology*, *56,* 24-34.

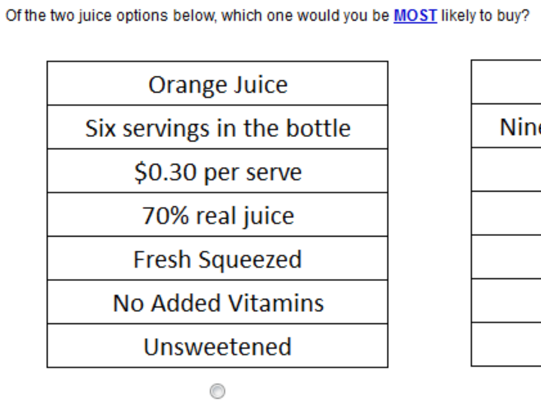
Mueller Loose, S., & Lockshin, L. (2013). Testing the robustness of best worst scaling for cross-national segmentation with different numbers of choice sets. *Food Quality and Preference*, *27*, 230-242.

Outwater, M., Spitz, G., Lobb, J., Campbell, M., Sana, B., Pendyala, R., & Woodford, W. (2011). Characteristics of premium transit services that affect mode choice. *Transportation*, *38,* 605-623.

Remaud, H., & Gillan, M. (2007). Understanding voters' expectations and behaviour: a best-worst analysis. *Australian and New Zealand Marketing Academy Conference*, Dunedin

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory and Cognition*, *21*, 546-556.

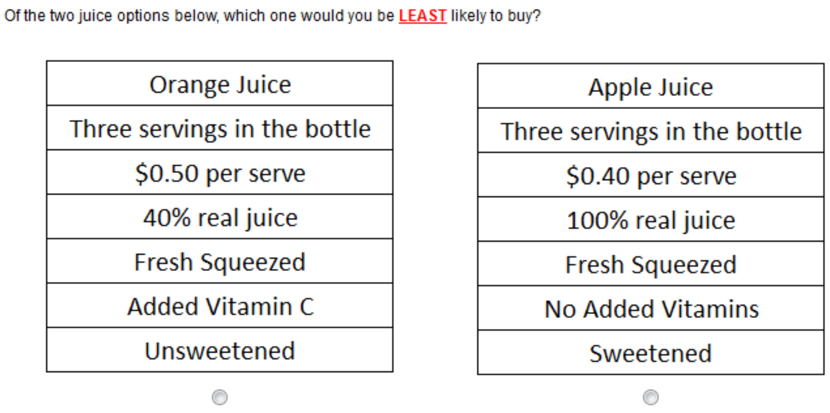
Appendices



Of the two juice options below, which one would you be **MOST** likely to buy?

Orange Juice	Orange Juice
Six servings in the bottle	Nine servings in the bottle
$0.30 per serve	$0.30 per serve
70% real juice	100% real juice
Fresh Squeezed	Fresh Squeezed
No Added Vitamins	Added Vitamin C
Unsweetened	Sweetened
○	○

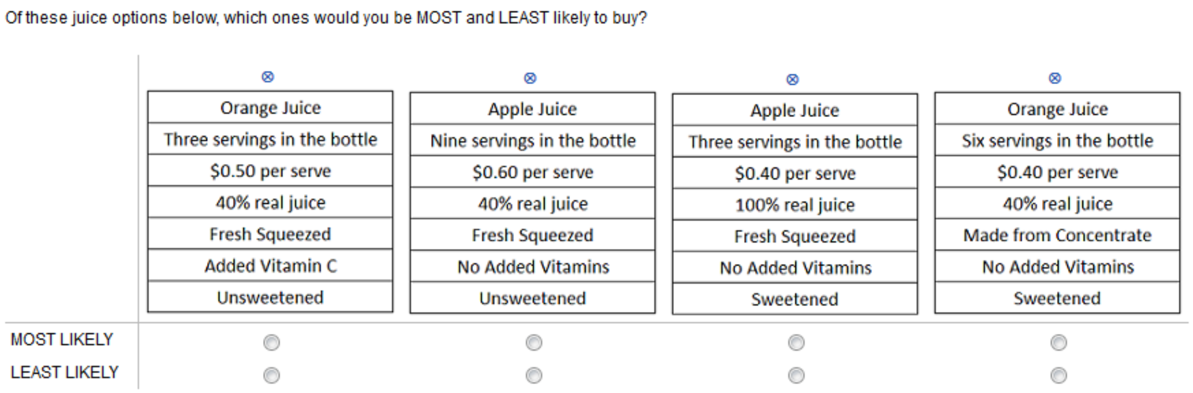
Figure A.1: Example of ‘Best’ condition choice set



Of the two juice options below, which one would you be **LEAST** likely to buy?

Orange Juice	Apple Juice
Three servings in the bottle	Three servings in the bottle
$0.50 per serve	$0.40 per serve
40% real juice	100% real juice
Fresh Squeezed	Fresh Squeezed
Added Vitamin C	No Added Vitamins
Unsweetened	Sweetened
○	○

Figure A.2: Example of ‘Worst’ condition choice set



Of these juice options below, which ones would you be MOST and LEAST likely to buy?

	⊗	⊗	⊗	⊗
	Orange Juice	Apple Juice	Apple Juice	Orange Juice
	Three servings in the bottle	Nine servings in the bottle	Three servings in the bottle	Six servings in the bottle
	$0.50 per serve	$0.60 per serve	$0.40 per serve	$0.40 per serve
	40% real juice	40% real juice	100% real juice	40% real juice
	Fresh Squeezed	Fresh Squeezed	Fresh Squeezed	Made from Concentrate
	Added Vitamin C	No Added Vitamins	No Added Vitamins	No Added Vitamins
	Unsweetened	Unsweetened	Sweetened	Sweetened
MOST LIKELY	○	○	○	○
LEAST LIKELY	○	○	○	○

Figure A.3: Example of ‘Best Worst’ condition choice set

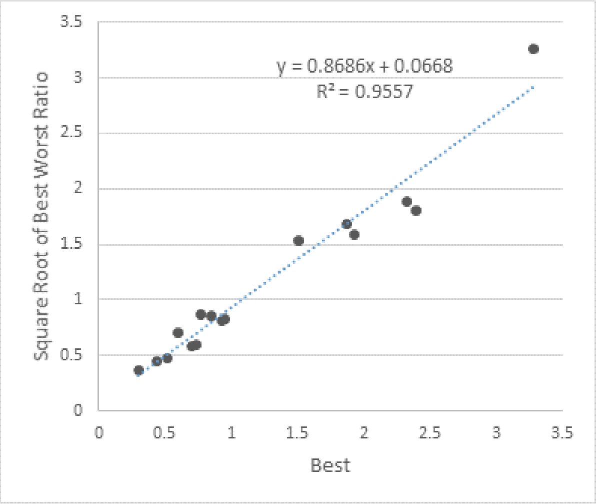


Figure B.1: Relationship between Best response and Square Root of Best-Worst Ratio for Juice Category

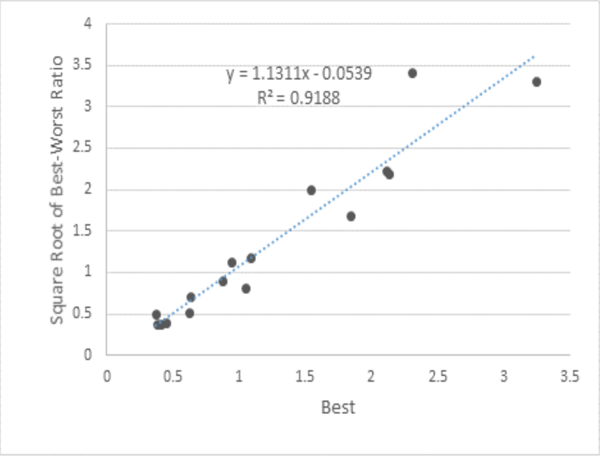


Figure B.2: Relationship between Best response and Square Root of Best-Worst Ratio for Pizza Restaurant Category

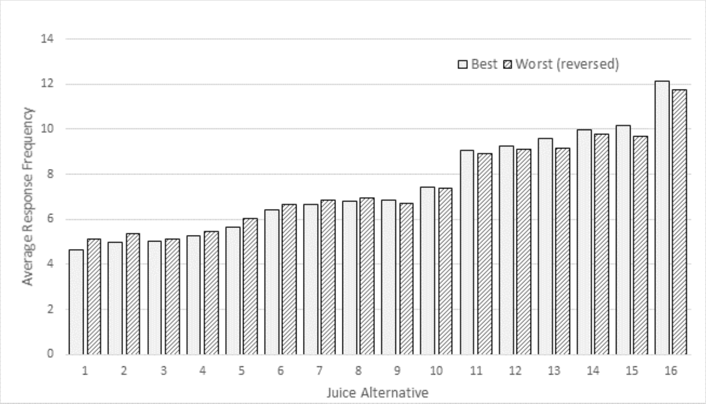


Figure C.1: Best and Worst (reversed) response frequencies in the Binary Experiments, Juice Category

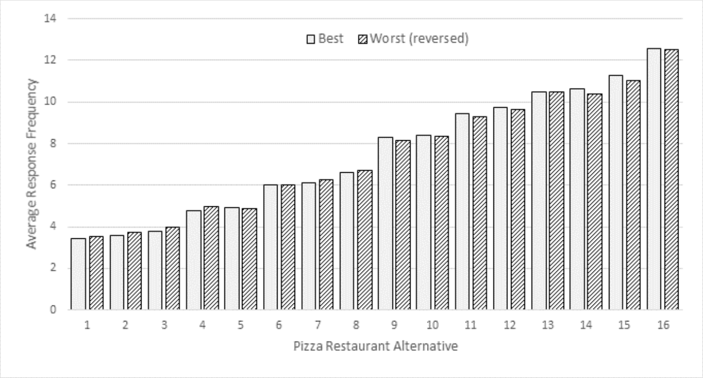


Figure C.2: Best and Worst (reversed) response frequencies in the Binary Experiments, Pizza Restaurant Category

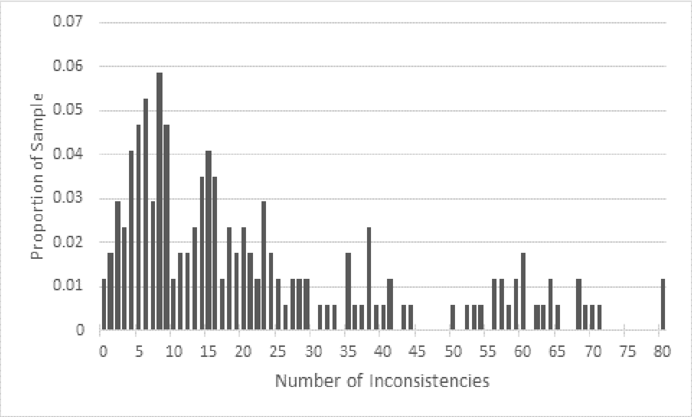


Figure D.1: Distribution of Inconsistencies between Best and Worst (reversed) responses in the Binary Experiments, Juice Category. Graph has been truncated on X axis with final column being ‘80 or more’.

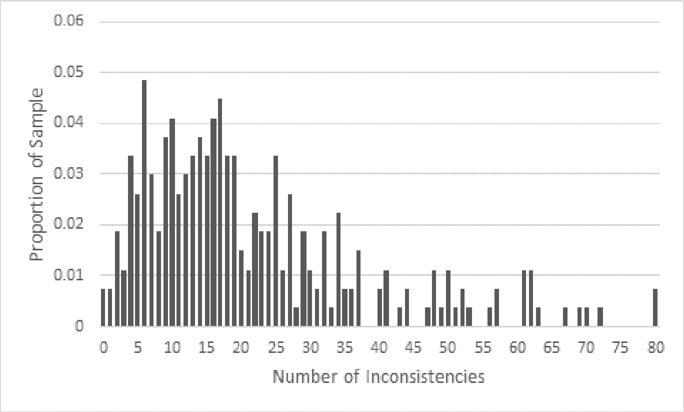


Figure D.2: Distribution of Inconsistencies between Best and Worst (reversed) responses in the Binary Experiments, Pizza Restaurant Category. Graph has been truncated on X axis with final column being ‘80 or more’.

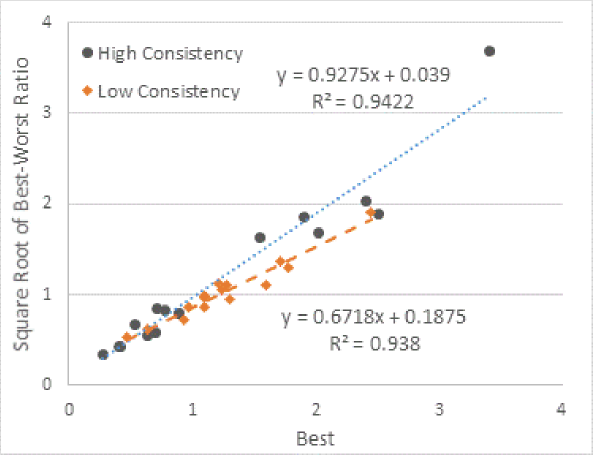


Figure E.1: Relationship between Best response and Square Root of Best-Worst Ratio in Juice Category

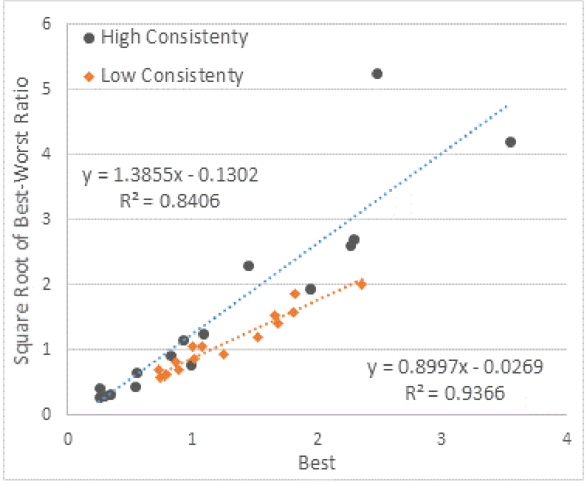


Figure E.2: Relationship between Best response and Square Root of Best-Worst Ratio in Pizza Restaurant Category