Australasian Marketing Journal 23 (2015) 227–234

Contents lists available at ScienceDirect

Australasian Marketing Journal

journal homepage: www.elsevier.com/locate/amj

Heterogeneity in the consistency of best–worst scale responses

Luke Greenacre *, Steven Dunn, Ana Mocanu

Ehrenberg-Bass Institute and School of Marketing, University of South Australia, SA 5000, Australia

ARTICLE INFO

Article history:
Received 19 September 2014
Revised 13 March 2015
Accepted 1 June 2015
Available online 17 June 2015

Keywords:
Choice
Experiment
Best–worst
Scale
Heterogeneity
Consistency

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. However, 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 provide 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.

CHINESE ABSTRACT

在选择实验中，最佳、最差级别是占主导地位的度量方法之一。使用中，它提供了有关人们喜好的实质性信息，且不会使选择任务过于漫长。然而，关于此方法的一种担忧是人们所选择的最佳方式，可能并不反映选择最差方式时的相同的偏好。对此种最佳、最差的回答之间的不一致性的研究，已经发现这是个不小的、持久性的问题。该研究进一步调查了这种不一致性，并发现它们在很大程度上可能归因于样品中的相对较小的一群人，他们并不像文献表述的会将最差的回答固定到最佳的回答上。事实上，在一个样品帐户中，25%的参与者占了所记录的50至60%的不一致回答。这个群体的存在，以及他们对最佳和最差的回答的不一致的数量的不成比例的影响，提供了强有力的证据，证明在人们如何始终如一地规定最佳、最差任务中的回答时，存在异质性。本文就如何在基于效用的选择模型中适应这一现象给出了建议，以更好的预测选择。

1. Introduction

The interest in Best–worst (BW) scaling has increased dramatically over the last few years (Flynn et al., 2007; Louviere et al., 2013; Marley and 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 and Gillan, 2007), policies and opinions (Finn and Louviere, 1992; Jones et al., 2013), means of transport (Outwater et al., 2011), medical treatments (Flynn et al., 2007), consumer products (Cohen, 2009; Louviere et al., 2013), or any other objects that people may naturally choose from. 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 et al., 2010; Louviere and 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 and Lockshin, 2013).

In a parallel literature, preference and response *heterogeneity* have been raised as a potential source of error in the analysis of BW data,

* Corresponding author. Tel.: +61 8 830 20017; fax: +61 8 8302 0442.
E-mail address: luke.greenacre@unisa.edu.au (L. Greenacre).

http://dx.doi.org/10.1016/j.ausmj.2015.06.002
1441-3582/

and in choice experiments in general (Cardell, 1997; Flynn et al., 2010; Hutchinson et al., 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 generalisation 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 and 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 amongst 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 recognises 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 and Louviere, 2005). Each alternative is a simple object that is expected to be holistically evaluated. For example this could simply be the choice amongst 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 et al., 2007). For example, 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 and 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 amongst 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 and Louviere, 2005; Marley and Pihlens, 2012; Marley et al., 2008). 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 and 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 to situations where BW data is used.

Testing has shown that there tends to be an agreement between people's best and worst responses (Mueller Loose and 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 and 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 and Pihlens, 2012). Some people may have weaker anchoring to the best, leading to the prompt for worst responses to activate a decision making process different from 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