Intern. J. of Research in Marketing 30 (2013) 292–303

Contents lists available at SciVerse ScienceDirect

Intern. J. of Research in Marketing

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

An introduction to the application of (case 1) best–worst scaling in marketing research☆

Jordan Louviere [a,*], Ian Lings [b], Towhidul Islam [c], Siegfried Gudergan [d], Terry Flynn [a]

[a] Centre for the Study of Choice (CenSoC), University of Technology, Sydney, PO Box 123 Broadway, NSW 2007, Australia
[b] School of Advertising, Marketing and Public Relations, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia
[c] Department of Marketing and Consumer Studies, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada
[d] The University of Newcastle, Social Sciences Building, University Drive, Callaghan, NSW 2308, Australia

ARTICLE INFO

Article history:
First received in 30, August 2008 and was under review for 9 months
Available online 2 April 2013

Area editor: Peter C. Verhoef

Keywords:
Best–worst scaling
Measurement
Preference
Choice

ABSTRACT

We review and discuss recent developments in best–worst scaling (BWS) that allow researchers to measure items or objects on measurement scales with known properties. We note that BWS has some distinct advantages compared with other measurement approaches, such as category rating scales or paired comparisons. We demonstrate how to use BWS to measure subjective quantities in two different empirical examples. One of these measures preferences for weekend getaways and requires comparing relatively few objects; a second measures academics' perceptions of the quality of academic marketing journals and requires comparing a significantly large set of objects. We conclude by discussing some limitations and future research opportunities related to BWS.

1. Introduction

Academics and practitioners in various disciplines often wish to measure an individual's strength of preference for (or level of agreement with) a number of objects (which can be statements or some other item of interest). A typical objective is to locate all the objects on a measurement scale with known mathematical properties to allow robust statistical comparisons of changes over time and/or differences among respondents. In practice, this can be challenging. For example, rating scales attempt to ensure that all individuals use the same numerical scale, but in practice, various idiosyncrasies in response styles have been found (Auger, Devinney, & Louviere, 2007). Such idiosyncrasies can arise from individuals using rating scales in different ways, from cultural differences and/or from verbal ambiguities with labels (Lee, Soutar, & Louviere, 2008). Furthermore, it has been observed that individuals tend not to discriminate between response categories when they are not asked to respond in ways that elicit tradeoffs or *relative* preferences for the objects being valued, such as asking people to rate the "importance" of several factors on a rating scale. That is, respondents do not have to trade off one factor against another, as evidence indicates that this often leads to minimal differences in mean ratings (e.g., Cohen & Neira 2003; Lee, Soutar, & Louviere, 2007).

An approach to dealing with such issues that has been growing in popularity in many fields is to avoid tasks that ask individuals to use numbers in favor of tasks that infer strength of preference (or other subjective, latent dimensions) from how often they choose one object over other, known objects. Such observed choice frequencies ensure that the derived numbers are on a known (choice frequency or probability) scale. However, some choice-based approaches, such as the method of paired comparisons, require large numbers of choice questions to estimate preferences for objects. Indeed, asking individuals to choose from all possible pairs of objects is not feasible in survey settings as the number of objects grows, a clear weakness of the method of paired comparisons.

The purpose of this paper is to introduce, discuss and illustrate a choice-based measurement approach that reconciles the need for question parsimony with the advantage of choice tasks that force individuals to make choices (as in real life). Prior work recognizes three choice-based measurement cases. In case 1 (the object case), individuals are asked to choose the best and worst (on some subjective scale) from a set of objects (e.g., Finn & Louviere, 1992). In case 2 (the profile case), individuals evaluate several profiles of objects described by combinations of attributes/features dictated by an underlying design; they "see" the profiles one at a time and choose the best and worst feature/attribute levels within each presented profile (e.g., Louviere, 1994). In case 3, individuals choose the best and the worst designed profiles (choice alternatives) from various choice sets dictated by an underlying design (e.g., Marley & Pihlens, 2012).

The purpose of this paper is to introduce, discuss and illustrate case 1. We focus on case 1 because it illustrates the fundamentals of choice-based measurement in general and what is known as "best–worst scaling" (BWS) in particular. BWS was introduced by Finn and Louviere (1992), and recent advances suggest that academics and practitioners would benefit from an updated discussion of its concepts and methods.

☆ This paper was prepared with the assistance of the Services Innovation Research Program, QUT Business School, Queensland University of Technology.
* Corresponding author. Tel.: +61 2 9514 9799; fax: +61 2 9514 9897.
E-mail address: jordan.louviere@uts.edu.au (J. Louviere).

0167-8116/$ – see front matter. Crown Copyright © 2013 Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ijresmar.2012.10.002

BWS is one way to avoid and overcome some of the limitations of rating-based and similar measurement methods used in marketing and in other fields. BWS case 1 typically allows one to obtain measures for each person (respondent) on a difference scale with known properties (Marley & Louviere, 2005). Cases 2 and 3 can be viewed as extensions of case 1 in which objects or items are represented as multi-dimensional choice objects (options). However, the fundamental ideas and principles from case 1 also apply to cases 2 and 3; thus, we focus on explaining case 1 in detail because this provides a foundation for understanding cases 2 and 3.

Accordingly, the objective of this paper is to provide an introduction for academics and practitioners on how to design, implement and analyze case 1 BWS studies. The case for such a paper is threefold. 1) As papers detailing the mathematical proofs of the main estimators used to implement such studies are highly technical and not easily understood by novices (Marley & Louviere, 2005; Marley, Flynn, & Louviere, 2008), there is a need for a more straightforward explanation to encourage applications. 2) Disciplines in which comprehensive 'how to' BWS discussions have been published have seen a proliferation of empirical studies (e.g., Flynn, 2010), suggesting that a tutorial paper should benefit marketing academics and practitioners. 3) Several methods for estimating the values of objects using underlying subjective scales have been proposed, but many of these, although easy to implement in a spreadsheet or generic statistical package, are not part of the typical 'toolbox' of methods used by academics and practitioners. Indeed, a 'user guide' paper detailing the BWS profile case (case 2) for health economists arose from requests at conferences to (among other things) 'see' what the data and regression models 'look' like (Flynn, Louviere, Peters, & Coast, 2007).

Accordingly, to provide a 'how to' BWS tutorial, this paper is organized as follows. First, we offer a conceptual framework and empirical justification for BWS. We then present two empirical studies. We emphasize how to set up, design and implement a BWS case 1 survey in practice, and how to analyze the associated results. Specifically, we present worked examples that illustrate how to use BWS for relatively small (six objects) as well as very large (72 objects) comparison sets. The paper ends with a discussion and conclusion section that recaps the major points of the paper, identifies some limitations and issues and suggests some potential future research directions.

2. A conceptual framework for BWS

BWS is underpinned by random utility theory (RUT), which also underlies discrete choice experiments used in marketing research and economics (McFadden, 1974; Thurstone, 1927). RUT assumes that an individual's relative preference for object A over object B is a function of the relative frequency with which A is chosen as better than, or preferred to, B. Thus, it requires individuals to make choices stochastically (with some error). Thurstone's (1927) paper proposed RUT and used it to motivate and develop the method of paired comparisons, where individuals choose the 'best' object from sets of two objects. Thurstone recognized that the theory requires individuals to make errors in their choices, thus allowing the model parameter estimates that we term 'scale values' to be derived. Scale values are measures of the locations of each object on an underlying subjective scale of interest. McFadden (1974) generalized Thurstone's RUT model to provide tractable, closed-form models that accommodate choices from sets of three or more objects. More formally, for the 'best' only case McFadden considered:

$$
\begin{aligned}
S_A &= V_A + \varepsilon_A \\
S_B &= V_B + \varepsilon_B \\
S_C &= V_C + \varepsilon_C \\
S_D &= V_D + \varepsilon_D.
\end{aligned}
$$

In the above, the true subjective scale value (S_k) of the kth object consists of two components, the observed value V_k, which is systematic (explainable), and the errors ε_k, which are random (unexplainable). The random component implies that one cannot predict the exact choice that a person will make, but only the probability that a person will choose each object offered (McFadden, 1974). This choice probability can be expressed as:

$$P(U = \text{best}|A, B, C, D) = P[(V_A + \varepsilon_A) > (V_k + \varepsilon_k)],$$

considering that all other options are available to be chosen in the comparison set. McFadden (1974) derived what is known as the conditional logit model by assuming that the errors are distributed as independent and identically distributed Type 1 Extreme Value. The choice probabilities for this model have the following closed form expression:

$$P(A = \text{best}|A, B, C, D) = \exp(V_A)/[\exp(V_A) + \exp(V_B) + \exp(V_C) + \exp(V_D)].$$

McFadden's framework relates choices from sets of multiple objects to an underlying latent scale value associated with each object, but until recently, little work was available to help researchers identify and implement reasonably good ways of *collecting* choice data from individuals to implement these models. An obvious exception, of course, is the method of paired comparisons, which has been extensively studied (e.g., David, 1988). Unfortunately, the method of paired comparisons poses inherent limitations in survey applications because the number of comparisons needed increases geometrically with the number of objects to be measured. Thus, paired comparisons can be practical for measuring a few objects (e.g., six objects require 15 pairs), but typically are not practical for larger numbers of objects (e.g., we later study 72 objects, which would require 2556 pairs).

One way to address the size limitation of paired comparisons is the multiple choice approach introduced by Louviere and Woodworth (1983) that relies only on 'best' choices. Although their discrete choice experiment ("DCE", also called "choice-based conjoint") approach is widely used, few researchers seem to appreciate that collecting only "first (or best) choices" provides minimal information for statistical estimation purposes. Thus, an approach that provides more statistical information than merely the first or best choice could be useful in many research applications.

BWS capitalizes on the fact that collecting 'worst' information, in a similar way to 'best' information, provides much more information. That is, BWS capitalizes on the idea that when individuals evaluate a set of three or more objects or items on a subjective scale, their choices of the top and bottom objects/items should be (all else equal) more reliable than choices of middle objects/items. Thus, BWS assumes that individuals make reliable and valid choices of the two most extreme objects/items in a set, consistent with the adaptation level theory (Helson, 1964). A key advantage of BWS is that it provides information about both the top ranked and bottom ranked items in a set. Taken together, these two choices provide much more information about the ranking of the choice options in each set. Only order information matters in choices; hence, asking for both top and bottom ranked choices provides much more information about the overall ranking of the objects than just the top choice.

More generally, BWS implies use of multiple comparison sets, with each set having at least three objects/items. In this respect, a BWS "experiment" is just another type of DCE, similar to the DCEs proposed by Louviere and Woodworth (1983). To wit, they proposed constructing comparison (choice) sets from 2^J fractional factorial designs (J = the number of objects/items). However, most BWS applications design choice (comparison) sets with balanced incomplete block designs (BIBDs), such as Lee et al. (2008). A BIBD is a type of experimental design in which each choice option appears equally often, and co-appears equally often with each other choice option. Unlike 2^J designs, BIBDs ensure that choice set sizes are always equal. A type of BIBD called a "Youden" design (e.g., Raghavarao, 1988) allows one to control for order by ensuring that each object appears in every order. In our experience, there is little difference in outcomes associated