Theory and Methods

Chapter 1

Introduction and overview of the book

This book is written for researchers and practitioners who have minimal prior knowledge of best-worst scaling (BWS). Many readers will have experience with discrete choice experiments (DCEs), and to avoid making this a two-volume book we take some knowledge of that field as a given. However, the purpose is explicitly *not* to build a subdiscipline that is accessible only to a small number of practitioners who are already experts in a highly technical field. On the contrary, we wish to show that BWS is accessible to the average applied practitioner and that in many cases it can be successfully implemented using spreadsheets rather than statistical programs. However, to do this we will refer to ongoing methodological and cutting-edge theoretical work and will draw on methods and techniques used across several disciplines. Thus, while the book should enable any moderately quantitative practitioner to run a BWS study, academics interested in any non-routine application should have (considerable) cross-disciplinary experience with the methods: many important insights are gained only by experience, and the days of the generalist health/environmental/transport economist doing a DCE or BWS study are (or should be) numbered. Nevertheless, we hope to encourage practitioners with knowledge about discrete choice methods to apply BWS, as many of the design and analysis techniques are simple, though not part of the typical analytical "toolbox" taught to academics and practitioners.

1.1 A brief history of BWS

Best-worst scaling was developed by one of the authors (Louviere) in 1987 at the University of Alberta. Louviere was curious about what could be done with data that resulted from asking people not only to report the "top" choice in each choice set but the "bottom" choice as well. He constructed a small design and made a choice task to go with it, and asked Tulin Erdem, a PhD student at Alberta, to "do" the task. Tulin "did" the task, and brought it back asking what it was for and how it worked. Louviere told her that he had no idea how it worked, but thought that it could be a useful addition to choice experiments because it provided extra choice information. Crucially, it provided information about less attractive choice options and much more information about the respondent's value (utility)

function. Louviere spent the next several years working on how to conceptualize such a task, and how to interpret and use the resulting data.

These efforts resulted in a paper with Adam Finn in the *Journal of Public Policy and Marketing* (Finn and Louviere, 1992), in which they showed how to apply BWS to typical public polling problems, illustrating the approach by quantifying public concern over food safety. Not only did they show that food safety was of little concern to their sample, but they also helped to avoid spending a significant amount of public funds on an advertising campaign to "convince" the public represented by the sample that their food supply was "safe." Louviere next wrote a section on another type of BWS in a chapter in Richard Bagozzi's *Advanced Methods of Marketing Research* (Louviere, 1994). This was followed by several working papers on BWS with Joffre Swait in the early 1990s. BWS languished during the 1990s despite attempts by combinations of Louviere, Swait and Donald Anderson to publish papers on the subject (all papers were rejected by academic marketing reviewers – a lesson in perseverance for young scholars).

Louviere was contacted by Emma McIntosh (Oxford University), who was interested in discrete choice experiments and thought that BWS was a promising way to approach several problems in health economics. This resulted in a talk by McIntosh and Louviere (2002) at a conference on applications of discrete choice experiments in health economics (a dental care application) and a chapter in McIntosh's PhD thesis.

Meanwhile, Louviere was motivated to persuade Tony Marley that BWS tasks were interesting and important. Much earlier Marley (1968) had proposed a complex probabilistic choice model that included judgments of "the superior alternative" and "the inferior alternative" in a choice set. However, Marley did not originally see the potential of those ideas for what became BWS. In time Marley signed on to provide formal theory linking the task to various cognitive processes that humans could use to "do" the task. Each such process potentially has different implications for how it can be formally represented (that is, "modeled"), as well as different implications for the mathematical properties of the resulting measures that one derives by applying the theory. This culminated in a paper on the theory of BWS and associated statistical properties in the *Journal of Mathematical Psychology* (Marley and Louviere, 2005).

Eventually, a number of marketing research practitioners were attracted to BWS because of its potential to avoid some of the problems associated with category rating scales, such as differences in the way individuals use them. One particular advocate, Steve Cohen, presented several papers on BWS at ESOMAR (originally the European Society for Opinion and Market Research) conferences (Cohen, 2003; Cohen and Neira, 2003), winning "best paper" awards for his efforts, which led to considerable interest in BWS by practitioners. Shortly afterwards Bryan Orme of Sawtooth Software teamed with Cohen to present a paper on the approach (Cohen and Orme, 2004); Sawtooth also produced commercial applications software to implement BWS in surveys. This led to more interest in BWS, which resulted in its widespread adoption and use by marketing researchers in many countries.

McIntosh's work led several other health economists, including Terry Flynn at Bristol (then in Sydney), to approach Louviere to collaborate on projects. These collaborations

produced several conference talks in health economics and papers, most notably a "how to do BWS" paper in the *Journal of Health Economics* (Flynn *et al*., 2007). Many other applications have followed in health economics, particularly since the National Institute of Health Research (NIHR) funding body in the United Kingdom publicly stipulated its use in research related to valuing social-care-related quality of life (Potoglou *et al*., 2011; NIHR, 2006).

Another parallel stream of interest and application arose in personality and values measurement. Julie Lee and Geoff Soutar of the University of Western Australia contacted Louviere for assistance and collaboration in applying BWS to Schwartz' list of values, resulting in invitations to speak at Schwartz' annual values conference in 2006, and two recent papers (Lee, Soutar and Louviere, 2007; 2008). The Australian Research Council has since funded a Discovery grant application by Lee, Soutar, Schwartz and Louviere to compare Schwartz' new refinements to the values categories and test them using BWS.

Interest in BWS also arose in food and wine research, with researchers from the Wine Marketing Institute at the University of South Australia collaborating with Louviere on several projects to measure the importance of various wine attributes, using BWS in sensory measurement. These collaborations culminated in a large grant from the Australian Grape, Wine and Brandy Research and Development Corporation to the University of South Australia to use choice experiments and BWS to model and predict demand for new wine styles in developing markets (Casini, Corsi and Goodman, 2009; Cohen, 2009; Goodman, 2009; Mueller and Rungie, 2009). Additionally, several researchers in Australia and elsewhere have begun studying ways to use BWS to obtain sensory measurements in food science and related areas. These collaborations and applications are only a few of many currently under way on BWS, and include only those currently known to us.

As a result of these collaborations and applications it became clear to us that there was a need for a book on BWS that brought together the theory and methods and illustrated their application in various case studies in one handy reference guide. Thus, the idea of this volume was born out of experience working with others on a diverse array of applications, seeing a clear need to bring together as much material on BWS as possible to help people get started in their learning of BWS theory and methods.

1.2 The plan of the book

The book is organized around three areas of BWS theory that we call (1) the object case (Case 1), (2) the profile case (Case 2) and (3) the multi-profile case (Case 3). Each case is presented in a separate chapter that includes basic theoretical results and discusses their meaning and implications. Formal statements of these theoretical results are given in Chapter 5, which also summarizes recent extensions of BWS that include measures of the time to make responses (viz, response time). Each case discusses the design of the relevant statistical experiments, before going on to describe the various methods of

analysis, using case studies. For Cases 1 and 2 there is also a discussion of how the theory can be implemented and applied, focusing on different processes that individuals can follow to provide best-worst data. (These processes are equally relevant to Case 3 but are omitted to avoid repetition.) Each theory and methods chapter is linked to three applications chapters written by collaborators or the authors, and used to illustrate the major ideas and themes in the theory and methods chapters. The "looking forward" Chapter 6 discusses limitations, theoretical and empirical research gaps and important problems that we would like to see resolved.

1.3 A non-technical introduction to BWS

Best-worst scaling is based on the idea that a person faces choices among collections ("sets") of three or more items or options, and can identify the best and the worst options in the collection. Here "best" and "worst" simply constitute a metaphor for any appropriate terms that define the extremes of a latent, subjective continuum. For example, one can think of a set of three or more people, with the extremes being "tallest" and "shortest," or a set of three or more weights, with the extremes being "heaviest" and "lightest," or a set of three or more brief biographical sketches of individuals, with the extremes being "most like to meet" and "least like to meet." Many more such examples could be provided. At this point, it is important merely to recognize that these simple principles defining the extremes of a collection (continuum) are common to all types of BWS. The three cases that will be described merely differ in terms of how complex the items or options under consideration are. Moreover, a best option and a worst option in a set of available items are not, in general, the same as an *acceptable* (as in "would purchase") option and an *unacceptable* (as in "would not purchase") option in a set of available items; we discuss this important distinction in Chapter 6.

1.3.1 The object case (Case 1)

This is the "classic" case of BWS that was developed by Louviere in the late 1980s (Finn and Louviere, 1992). In this case a researcher is interested in measuring a set of objects, items, statements, people, pictures, product features, brands, towns, countries, environmental settings, health equity and efficiency issues in priority setting, public policy issues, etc. on an underlying, latent, subjective scale. For example, one may want to measure individuals' perceptions of: product feature importance, brand quality, public issue priorities, the attractiveness of persons, degrees of agreement with statements, the scenic beauty of towns or environmental settings, or the priority that certain types of people, such as smokers, should be given in setting health policies for treatments.

Thus, the object case requires one to have a "list" of items, objects, people, brands, etc. that one wants to measure. This list is exogenous to BWS projects, but integral to them. In any event, if one has a list of objects (we now refer to "objects" in this section, but it should

I think that this is the best Airline (☑ one)	Airline	I think that this is the worst Airline (☑ one)
☐	American	☐
☐	United	☑
☑	Qantas	☐
☐	Delta	☐

Figure 1.1 Example choice set containing four airlines

be clear that "objects" refers to any list), the objective is to measure each object on one or more underlying, latent subjective scales. To do this one must complete a series of steps consistent with the theory discussed in Chapter 2.

The objects, items, people, statements, etc. are systematically organized into subsets of three or more, and a sample of people evaluate each of the subsets and make best and worst (most and least, etc.) choices in each subset. Figure 1.1 presents a hypothetical choice set containing four airlines that one may consider for a trans-pacific flight.

The principles underlying the analysis of the best-worst choice data are similar to those in a discrete choice experiment, and the theoretical framework common to both is random utility theory (RUT). RUT assumes that people make errors, but when choosing repeatedly their choice frequencies give an indication of how much they value the items under consideration (Thurstone, 1927). Thus, how often item A is picked over item B gives an indication of how much item A is preferred to item B. So, how often a respondent picks Qantas over United on the trans-pacific airline route provides an estimate of how much he/she prefers Qantas to United; in particular, with best-worst choice, this information is obtained from how often the respondent selects Qantas (respectively, United) as best and/or United as worst in the presented choice sets.

A key issue for DCEs and BWS concerns the composition of the choice sets: what subsets of items (in this case airlines) should be presented to respondents? Chapter 2 will detail the principles of experimental design used to achieve this in the object case of BWS; at this point it is important merely to understand why asking for least preferred as well as most preferred is useful. Using the airline example, the researcher may notice that for the "most preferred" choices (for example) a particular respondent picks one of the Star Alliance airlines only when there are no airlines available from the OneWorld airline partnership. However, those data will provide little information as to *which* Star Alliance airline is *least preferred* by a respondent. Presenting additional choice sets containing different combinations of Star Alliance members is an inefficient way of inferring the least preferred airline when a researcher can obtain that information from asking additional "least preferred" choices in choice sets that a respondent *has already considered*.

Asking for two pieces of information per choice set raises the question of how and if one should combine them into a single outcome variable. Chapters 2 to 4 will explain this issue, but at this point it is sufficient to note that, under very mild assumptions about how a person makes choices, there are a variety of mathematically acceptable ways that the best and

worst data can be combined; a single outcome variable can draw on the strength of the best data to make inferences at the "top" of the utility function and the worst data to make inferences at the "bottom" of the utility function. Chapter 5 presents further formal material on the aggregation of best and worst data, and Chapter 6 discusses when such aggregation makes sense.

1.3.2 The profile case (Case 2)

Here "best" and "worst" choices refer to attribute levels described/displayed as "profiles." "Profile" is a commonly used term in the conjoint analysis literature (Louviere, 1988b) that refers to a combination of attribute levels. Specifically, a profile is a single treatment combination from an experimental design. In other words, a product, a person, a holiday place, a transport mode, a job, etc. can be described by an underlying, basic set of attributes (features, factors, characteristics, dimensions, etc.) that pertain to and describe all specific members of a generic class of products, persons, holiday places, etc. Each attribute is represented by two or more values called "levels" that typically are chosen to span the entire domain of the class. For example, if the price is an attribute of a certain type of product, the levels used to represent and vary the price should span the range of recent prices or a range of prices expected to occur in a future period of interest. Some attributes are quantitative or numerical and are (more or less) continuous; other attributes, such as a person's gender, are discrete types of attributes (they have mutually exclusive, discrete values). Each attribute has its own unique type and suitable number of levels.

The levels of three or more attributes (features) are systematically combined into a *profile* (namely a description of a product, person, place, etc.). Each profile (combination of attribute levels) can be viewed as a subset of choice options, with a choice option in this case being one of the presented attribute levels. As in the object case, a sample of people evaluate each of the profiles (subsets) and make best and worst (most and least, etc.) choices of the attribute levels that describe each profile.

We illustrate Case 2 with a more complex version of the airline example given earlier. In fact, this example is based on a real online choice experiment that was administered by the Institute for Choice at the University of South Australia. The experiment focuses on profiles of airline tickets for long-haul flights (such as Boston to Seattle, Sydney to Perth, Haikou to Burgin, Moscow to Vladivostok, Montreal to Vancouver, etc.). Each ticket profile is described by six attributes and associated levels (in parentheses): the round-trip airfare ($350, $450, $550, $650), the total flying time (3hrs, 4hrs, 5hrs, 6hrs), the airline name (American, Delta, Northwest, United), frequent flyer points (no, yes), the number of stops en route (0, 1) and whether there are free drinks en route (no, yes). Figure 1.2 presents a hypothetical Case 2 choice set containing a particular ticket (profile) described by the six attributes.

Again, respondents are asked to respond to a number of such profiles (choice sets of airline ticket attribute levels). All possible combinations of these attributes and levels

I think that this feature is the most attractive (☑ one)	Ticket option features	Specific details of Ticket Option	I think that this feature is the least attractive (☑ one)
☐	Airfare	$650	☐
☐	Travel time	6 hours	☑
☑	Airline	United Airlines	☐
☐	Freq flyer pts	Yes	☐
☐	No. of stops	None (direct)	☐
☐	Free drinks	No	☐

Figure 1.2 Hypothetical choice set for a Case 2 study

(tickets) can be represented as a $2 \times 2 \times 2 \times 4 \times 4 \times 4$ (or $2^3 \times 4^3$) factorial (512 combinations or profiles). Chapter 3 will describe how to select a suitable experimental design, as in many cases the researcher may be unable to administer all 512 profiles (and certainly not all 512 to each respondent).

Once each respondent has answered the choice sets the researcher must analyze the choice data. As for Case 1, analysis is usually conducted within a random utility theory framework. How often a feature (attribute level) is picked as best provides an indication of how much it is liked, and how often a feature is picked as worst provides an indication of how much it is disliked. Again, the worst data provide much better estimates (in terms of statistical precision) of the unattractive features of an airline ticket. For example, the best choice data from a respondent who never picked $650 airfare or six-hour flying time as best tell us nothing about which of those two features is less attractive, whereas the worst data almost certainly do. As for Case 1, the best and worst data are usually pooled to draw strengths from each, and Chapter 3 describes how this pooling is achieved.

Much more detail is provided in Chapter 3 for the following topics: (1) different ways to design profile experiments, statistical properties of these ways to design experiments, and the pros and cons of each; (2) how to translate designs into profiles, how to block profiles into versions to obtain more statistical information, and the pros and cons of each; (3) how to lay out and administer best-worst choice experiments for profiles, how to ask best-worst questions (including repeated best-worst questions), and the pros and cons of each; and (4) how to do basic and more sophisticated analyses of best-worst choice data, how to derive measurement scales for attribute levels, and test if the implied assumptions/properties underlying the models hold for empirical best-worst choice data, and the pros and cons of each.

1.3.3 The multi-profile case (Case 3)

The third case is associated with classical discrete choice experiments – that is, a person is offered a sequence of choice sets, with each choice set having three or more profiles (Louviere, Hensher and Swait, 2000; Hensher, Rose and Greene, 2005). The person's

Ticket option features	Ticket 1	Ticket 2	Ticket 3
Airfare	$650	$450	$550
Travel time	3 hours	5 hours	4 hours
Airline	United Airlines	Delta Airlines	American Airlines
Freq flyer pts	Yes	No	Yes
No. of stops	None (direct)	1	None (direct)
Free drinks	No	Yes	No
I'm most likely to choose (☑ one)	☑	☐	☐
I'm least likely to choose (☑ one)	☐	☐	☑

Figure 1.3 Example Case 3 choice set for airlines

task is to choose two profiles that are, respectively, the best and the worst (most, least preferred; most, least attractive; most, least like them; etc.).

As in the BWS object and profile cases, we have to identify a relevant "list" of things to be measured. In the multi-profile case, the list consists of attributes or features of options to be offered to people. Returning to the airline example, these might be the same six attributes used in the profile case: round-trip airfare, total flying time, airline name, frequent flyer points, number of stops en route, and free drinks en route. In the case of mobile (cell) phones, the features might be price, brand, camera/megapixels, Bluetooth capability, international roaming capability, GPS, etc.; in the case of delivered pizza products, the features might be brand, price, type of crust, number of toppings, delivery time, etc.; in the case of holiday destinations, the features might be type of environmental setting, flying or driving time, typical daytime high temperatures, range of activities, total cost, etc.

Once a list of features is determined, one must assign values or levels to each feature to represent a range of relevant possible variations, as in the profile case. Thus, for the airline example, if the same levels for each of the six attributes are considered, then the full factorial of airline tickets is still 512 profiles. The difference between this case and the profile case is that, instead of experimental participants evaluating one profile at a time, they evaluate three or more at a time. Instead of making choices *within* a profile (which feature of this single ticket is most/least attractive), they make choices *between* whole profiles (which ticket is most/least attractive), as they would do in a traditional DCE. We now return to the airline example to illustrate this case with an example choice set, as shown in Figure 1.3. Note, again, that the best-worst choices are conditional on the presented set of profiles (options) and some, all, or none, of those options may be acceptable (say, in the sense of possible purchase); we discuss this fact in Chapter 6.

As before, one must select or construct an experimental design to create the choice sets. This is more complex than Case 2: as well as deciding if one wants to use only a subset of the full factorial, one must decide how (in this case) to construct the sets of three tickets. We discuss the design of multi-profile choice experiments, or discrete choice experiments, in Chapter 4.