Best-Worst scaling...reflections on presentation, analysis, and lessons learnt from case 3 BWS experiments

Best-Worst scaling...reflections on presentation, analysis, and lessons learnt from case 3 BWS experiments

Jannie Mia Adamsen, Griffith University
Associate Professor Sharyn Rundle-Thiele, Griffith University
Dr Jennifer A. Whitty, Griffith University

ABSTRACT

Surveys based on Likert scales and similar ratings-based scales continue to dominate market research practice despite their many and well-documented limitations. Key issues of concern for Likert scales include over- or under-reporting depending on the context, and variation in responses based on cultural background. Alternatives exist to overcome the inherent weaknesses of these scales. This paper reflects on the Best Worst (BW) Scaling method that we have recently used in eight online studies. In these studies, we employed a novel pictorial approach to capture product preferences for over 3,600 respondents. One case 3 BW experiment investigating consumer preferences for organic apples is featured and evaluated using two approaches. The first analysis treats the data as a case 1 BW experiment to outline the simplicity of case 1 analysis. Case 3 BW analysis involving multinomial logit and latent class modelling is then illustrated to highlight the rich preference insights that can be obtained from case 3 BW studies. We look at some of the drawbacks of the BW case 3 approach, including design and data processing issues, weighted against the overall positive feedback received from respondents. Overall, we do believe the BWS method has a significant potential to improve predictability in market research – the response rate and positive participant feedback speaks for itself.

Keywords: Best-Worst Scaling, choice, Likert scales, online surveys, pictorial, apples, organic, case 3 Best Worst

Corresponding Author: Jannie Mia Adamsen, Department of Marketing, Griffith University
170 Kessels Rd, Nathan QLD 4111 Australia Email: j.adamsen@griffith.edu.au Phone: +61 7 3735 3879

Associate Professor Sharyn Rundle-Thiele, Department of Marketing, Griffith University
170 Kessels Rd, Nathan QLD 4111 Australia Email: s.rundle-thiele@griffith.edu.au Phone: +61 7 3735 3879

Dr Jennifer A. Whitty, Senior Lecturer, School of Medicine, Griffith Health Institute, Griffith University
Logan Campus, Logan QLD 4131 Australia Email: j.whitty@griffith.edu.au Phone: +61 7 3382 1486

1. INTRODUCTION

Current pressures (e.g. return on investment and declining response rates) require market research professionals to keep abreast of the latest developments in tools and techniques. In addition to employing the traditional repertoire of quantitative and qualitative market research techniques, today's market research professionals (New York AMA Communication Services Inc., 2013) are employing a diverse range of technologies that can analyse available information (web crawl, Google analytics, content analysis of consumer blogs) along with alternate data collection tools (e.g. Internet chat rooms, social media, blogs). Some marketing problems, e.g. new to market product formulations, will continue to rely on traditional research methods to provide the information needed to formulate a product solution to better meet the needs of the market. It is important for market researchers to understand the limitations of traditional quantitative research methods to ensure their clients receive maximum return on investment.

Many quantitative surveys employ simple rating scales, such as Likert scales to establish the extent to which a customer agrees/disagrees with a statement or likes/dislikes a potential new product. Likert scales have been widely used to measure a wide variety of issues including customer satisfaction, customer loyalty, trust, and service quality perceptions.. Considerable evidence exists suggesting Likert scales are not sufficiently reliable (see for example Chrzan & Skrapits, 1996; Cohen & Markowitz, 2002; Cohen & Neira, 2003; Louviere, Swait, & Anderson, 1995). This paper considers issues surrounding simple rating scales, such as Likert scales, and then continues by introducing a choice modelling technique, Best-Worst Scaling (BWS) that overcomes many of the reliability issues that are inherent with simple rating scales. Application of a novel pictorial approach aimed at reducing respondent burden is described. An overview of different analysis approaches is presented with reflections on the benefits and limitations of the specific method (BW case 3). Overall this paper

reflects on both researcher experience gained and respondent feedback received during the conduct of eight pictorial BWS studies conducted over a two year period (2009 through to 2011).

2. BASIC RATING SCALES – TOO BASIC?

Researchers have sought to understand the inherent limitations associated with simple rating based scales for over 60 years signifying that these are subject to many limitations, including, but not limited to, socially desirable responding, acquiescence bias, hypothetical bias and scalar equivalence. These limitations impact on a simple rating scales' ability to accurately measure across target populations. Given the known limitations of simple rating scales' it seems surprising that rating scales remain so widely used in market research. We now consider some of the criticisms directed towards simple rating scales.

Purchase intentions captured through simple rating scales, such as the Likert scale, have long been criticised for their inability to accurately predict customer behaviour (examples include Holdershaw, Gendall, & Wright, 2011 in a blood donation context and; Lockie, Lyons, Lawrence, & Grice, 2004 in an organic food context). As Lockie et al. (2004, p. 141) state *"if National Food Choice Survey estimates that organic food has captured about 1% of the Australian food market are accurate, then it would appear that a degree of over-stating has occurred"*. This issue is expanded further in a blood donation context by Barkworth, Hibbert, Horne &Tagg (2002) who note the problem with intentions captured through simple rating based scales are that while they are correlated with behaviour, they are not the same as actual donation behaviour. Lusk, McLaughlin & Jaeger (2007, p. 41) conclude: *"there is considerable evidence that inconsistencies often exist between what people say they will do and what they actually do"* and Holdershaw et al. (2011) suggest that researchers may now need to turn to other methods to predict behaviour.

The mismatch between what survey respondents indicate they will do and what they actually do may be a direct result of biases, including social desirable responding (SDR) and acquiescence bias. The idea underpinning SDR is that most respondents have a tendency to answer researchers' questions in ways that make themselves look good according to current cultural terms causing respondents to over-report or under-report, depending on the situation (Baumgartner & Steenkamp, 2005; Mick, 1996). SDR is particularly evident in situations that are socially sensitive, e.g. alcohol consumption or for topics where strong public opinion is present, e.g. sustainability. Examples of socially desirable responding abound. Consider one of the early studies by Zinkhan and Carlson (1995), who outlined an example of survey respondents' eagerness to describe themselves as recyclers which were incongruent with recycling rates at the time. Recently Rundle-Thiele (2009) provides an example in an alcohol context where survey respondents over reported alcohol free days (no alcoholic beverages consumed in the previous 24 hours) and under reported the actual amount of alcohol consumed in the previous 24 hour period. Researchers (examples include Bentler, Jackson, & Messick, 1971; McClendon, 1991; Welkenhuysen-Gybels, Billiet, & Cambré, 2003) have argued there is considerable evidence the Likert-type format is susceptible to SDR.

Scalar, or score, equivalence (Steenkamp & Hofstede, 2002) refers to the fact that rating scores obtained may not be directly comparable across countries or groups of consumers based on cultural heritage (Walley, Parsons, & Bland, 1999). Indeed, as noted by Steenkamp and Hofstede (2002, p. 202): *"It is worrisome to note that score equivalence has not received much attention in international segmentation research…We believe that lack of attention to score equivalence is one of the reasons why international segmentation studies often report a heavy country influence"*. Many differences observed using simple rating scales may occur due to an individual's interpretation of the scale's meaning. There is ample evidence (Baumgartner & Steenkamp, 2001; Chen, Lee, & Stevenson, 1995; Steenkamp & Baumgartner, 1998; Steenkamp, Hofstede, & Wedel, 1999) that countries differ in their response styles. Further, there is additional evidence suggesting that rating scales are susceptible to under- or over-reporting depending on the situation (Bentler, et al., 1971; McClendon, 1991; Welkenhuysen-Gybels, et al., 2003). The BWS method may minimize scalar equivalence as survey respondents are simply asked to identify the "best" and "worst" option in a fixed choice scenario, hence culturally dependent differences in the use of rating scales are minimised (Auger, Devinney, & Louviere, 2007).

Simple rating scale items (e.g. Likert scales) have often been shown to be susceptible to an acquiescent response bias (Billiet & McClendon, 2000; McClendon, 1991; Watson, 1992). Acquiescence is commonly defined as the tendency of respondents to show greater acquiescence (tendency to agree) rather than disacquiescence (tendency to disagree) with items irrespective of the content of that item (Baumgartner & Steenkamp, 2001; Billiet & McClendon, 2000; Rossi, Gilula, & Allenby, 2001; Watson, 1992). Related is the

notion of Extreme Response Style (ERS), which implies the tendency of respondents to disproportionately use the extreme categories in rating scales (Baumgartner & Steenkamp, 2001). Acquiescence bias is particularly prominent in new product development (NPD) research, where respondents give positive connotations to most new ideas (Zikmund, Ward, Lowe, & Winzar, 2007). The authors acknowledge that alternatives, such as Item Specific (IS) Response options (see Saris, Revilla, Krosnick & Schaeffer, 2010), exist to overcome acquiescence bias. However, we caution that IS response options then serve to limit the analytical techniques as the data is categorical. Moreover, IS responses remain incapable of dealing with multi-attribute options, which have been shown to be more realistic of real marketplace choices. Finally, while IS response options may overcome some of the limitations, especially in regards to acquiescence, of our standard Likert scales they still allow a respondent to agree/rate positively or disagree/rate negatively with all options, which is one of the major downfalls of rating scales in market research, especially for NPD and Willingness-To-Pay type research. It is important to note that in stated choice preference literature, these issues are often referred to as hypothetical bias, where respondents report a willingness to pay (WTP) that exceeds what they actually pay using their own money in laboratory or field experiments (Loomis, 2011, p. 363).

Recent research (Weijters, Cabooter, & Schillewaert, 2010) has furthermore found that the number of response categories, for example 5- or 7-point Likert scales, as well as whether all levels have been labelled or not has a significant influence on the overall results and conclusions. A fully labelled scale format was found to lead to higher acquiescence bias and lower ERS; the first due to the clarity provided by labels which strengthens the effect of positivity bias, the latter due to the increased salience and attractiveness of the intermediate options. Weijters et al. (2010, pp. 244-245) concludes that *"data obtained with different formats are not comparable, and interpretations of Likert data are always relative: the probability that respondents agree with an item depends on how such agreement can be expressed...The practice of reporting survey results by means of percentages of respondents who agree with a statement ("top two boxes" or "top three boxes") has to be treated with great caution"*.

Taken together, there is considerable evidence that simple rating scales are susceptible to many biases and equivalence issues which questions conclusions drawn from studies employing these measures (examples include Chrzan & Skrapits, 1996; Cohen & Markowitz, 2002; Cohen & Neira, 2003; Louviere, Swait, & Anderson, 1995). It is suggested, that BWS is in fact capable of minimizing biases such as SDR, acquiescence and scalar inequivalence and that BWS may produce better in-market predictions, even across different cultures, due to a unidimensional scale (Auger, Devinney, & Louviere, 2007; Goodman, Lockshin, & Cohen, 2006). Furthermore, recent research concludes that conjoint techniques, such as BWS, provides less dramatic hypothetical bias (Hensher, 2010; Loomis, 2011; Murphy, Allen, Stevens, & Weatherhead, 2005; Özdemir, Johnson, & Hauber, 2009). Taken together, these results suggest alternative methodological techniques including BWS may, indeed, assist market researchers to minimise bias.

With the limitations of simple rating scales in mind we have been using BWS in recent market research projects to establish organic food and baby care product preferences. We now continue by introducing BWS formally.

3. CONJOINT ANALYSIS

The following section will provide an overview of the different conjoint techniques: from the birth of rating- and ranking-based conjoint (classic conjoint approaches), through to choice-based conjoint and finally the method employed: Best-Worst Scaling (BWS). The overarching idea behind all of these techniques, however, is the same: they overcome the majority of the limitations of rating scales addressed above (American Marketing Association, 1992; Walley, et al., 1999), in particular:

- *Conjoint analysis is based on the assumption that purchase decisions are not made on a single factor but are based on several factors, or attributes, which are considered conjointly*
- *Traditional research techniques which aim to establish the importance of various product attributes invariably results in most attributes being classed as 'extremely important'*

Conjoint analysis does not simply ask respondents which attributes are important or which attributes they prefer, rather, it forces them to make trade-offs between the attributes and/or products. Preferences are then revealed through a series of rating (real score), ranking (implicit score) or trade-off decisions and it is argued that conjoint in this way overcomes the problem of respondents saying one thing and actually doing another, thereby providing results with higher validity and reliability as well as being more useful for marketing managers overall (Walley, et al., 1999). Results based on conjoint techniques are also said to provide better prediction and forecasting

models for consumer behaviour, especially for multi-attribute products or services (Green & Srinivasan, 1978; Wittink & Walsh, 1988).

3.1 The birth of conjoint

Back in the 1970's conjoint analysis as a technique gained foothold as a way of capturing consumer trade-offs among multiattribute products and service. A comprehensive review was written by Green and Srinivasan (1978), and conjoint analysis has since been used extensively in e.g. transportation (see for example: Rose, Hensher, Greene, & Washington, 2011) and environmental evaluation (Alriksson & Öberg, 2008). It is even argued that conjoint analysis *'is considered among the major contributions of marketing science to marketing practice"* (Netzeret al., 2008, p. 338). The overarching reason behind this popularity is that conjoint analysis in its basic form is decompositional, hence a consumers' preference for a given product or service can be decomposed into preference scores (often referred to as marginal utilities) for each component or attribute level (Cattin & Wittink, 1982, p. 46). Additionally, it is argued that by simulating real marketplace situations, conjoint analysis realistically models day-to-day consumer decisions and has a reasonable ability to predict consumer behaviour for multi-attribute products or services (Green & Srinivasan, 1978). It is important to understand that conjoint analysis is based on two underlying assumptions (Jaeger, Hedderley, & MacFie, 2001):

1. Consumer behaviour and subsequent choice is based on utility maximisation
2. Any product or service is basically a bundle of attributes from which consumers gain value

Conjoint analysis in its original form is based on ratings and rankings of alternatives. For a comprehensive overview of conjoint analysis and its early commercial application refer to Cattin and Wittink (1982).

3.2 Discrete Choice Experiments

More recently choice-based conjoint, also referred to as a Discrete Choice Experiment (DCE), has gained popularity. The DCE is rooted in Random Utility Theory, a well-tested theory of human decision making hypothesised by Thurstone (1927) and generalised by McFadden (1974). It was established that arbitrary ratings to model choice are not necessary; how often one option is chosen over others is enough (Zikmund, Ward, Lowe, Winzar, & Babin, 2011, p. 528). The general technique to reveal the subsequent preferences is based on regression methods such as multinomial logit (MNL), which also provides information about the influence of one attribute over others.

The overarching reason why the DCE has become increasingly popular is that it addresses one of the limitations of standard conjoint: in evaluations of products or services a high score in itself is no guarantee that a product will be chosen (Zikmund, et al., 2011, p. 528). In other words, in the DCE respondents have to make a choice or decision; depending on the application of the method (Jaeger, et al., 2001). Whether results are different between the original rating based conjoint and DCE is subject to debate (Huber, Wittink, Johnson, & Miller, 1992; Karniouchina, Moore, van der Rhee, & Verma, 2009; Moore, 2004; Oliphant, Eagle, Louviere, & Anderson, 1992). However, it would be fair to assume that a choice task would likely be more reflective of actual marketplace behaviour than a rating exercise, simply because making choices is what we all do every day (Huber, et al., 1992, p. 2). It is argued that respondents find it *'simpler and easier to compare objects and to simply select the one they would buy"* (Zikmund, et al., 2011, p. 532).

The overall question is then why DCE's are not extensively used in market research? The simple answer is that it is complicated to analyse the results (via MNL) and it is primarily left to expert consultancies to undertake. The simplest way to analyse DCE data is to purchase capable software, but that does not come cheap. The most popular packages are Sawtooth Software (US based) and NLogit (based on Econometric Software in the US but primarily designed by David Hensher and John Rose in Sydney). This leads us to what has been called the *"middle ground"* BWS (Flynn, et al., 2007).

3.3 Best-Worst Scaling – an overview

The central idea behind BW scaling is that participants are presented with a limited set of a larger number of objects/products/concepts, and are required to make two choices: the best (or most attractive, most useful, etc.) and the worst (or least attractive, least useful, etc.) (Zikmund et al., 2007). BWS originates from the same random utility framework that underpins other DCE and ranking studies and is generally seen as a good compromise between the two: more information is obtained with BWS than DCE, yet less burden is placed on the respondent than a full ranking of all choice options (Flynn, 2010, p. 259). Further, respondents are not asked to report how much they prefer different alternatives as with traditional numerical rating conjoint studies, they are merely asked to identify which of a number of options they prefer and which they do not (James & Burton, 2003).

The implication is that no participants are permitted to like or dislike all alternatives, as participants are forced to choose one most and one least preferred option in every scenario. A number of different object

sets are presented, to gather sufficient information about relative preferences from each respondent (Auger, Devinney, & Louviere, 2007; Cohen & Neira, 2003). Although there have been published papers utilising BWS over the past 15 years (Finn & Louviere, 1992), the formal statistical and measurement properties were proven only recently (Marley & Louviere, 2005).

Recently, papers have emerged discussing different types of BWS, namely the object case (case 1), the profile case (case 2) and the multiprofile case (case 3) (Flynn, 2010). In practice case 1 and 3 have been applied in the marketing field (see for example Steve Goodman, 2009 (case 1); Mueller, Lockshin, Saltman, & Blanford, 2010 (case 3)) and case 2 in health economics (see for example Potoglou et al., 2011). Case 2 has primarily been used in valuation studies concerned with general population preferences for e.g. quality of life attributes, and it is generally acknowledged that this approach is most appropriate when respondents have no experience with choice-making in the particular area of investigation, as profiles are presented one at a time (contrary to choice sets of two or more) (Potoglou, et al., 2011, p. 4). We will briefly introduce case 1 and case 3, however, the main focus and associated examples of analyses will be based on a case 3 BWS study to illustrate Case 3 Best Worst to its full extent. The main difference between case 1 and case 3 is that for a case 1 BWS study objects (which might be an attribute or profile) are simply presented as stand-alone measures and evaluated as such, whereas in case 3 studies attributes are bundled into a product/service. Figure 1 shows a simplified example of each case focusing on apples, which will be the product considered later in this paper. Please refer to Flynn (2010) for a detailed explanation of the three different types of BWS, examples and associated analyses.

Figure 1: Example of case 1 and case 3 BWS(adapted from Flynn, 2010):

BWS Case 1:
Please consider you are out shopping and want to buy apples.
Tick which attribute is most and least important to you.

Best/Most		Worst/Least
	Production method	
	Price	
	Packaging	
	Appearance	

BWS Case 3:
Please consider you are out shopping and want to buy apples.
Tick which apple product would be the best and worst for you.

Apple 1	Apple 2	Apple 3
Organic AU$8.99/kg Packaged B-grade	Conventional AU$6.99/kg Packaged A-grade	Organic AU$7.99/kg Loose-weight A-grade
Best ☐	Best ☐	Best ☐
Worst ☐	Worst ☐	Worst ☐

In regards to case 3 BWS studies, a major strength is that choices are presented in context and explicitly highlight the trade-offs that often have to be made during the decision-making task. In this sense, results are likely to be more reliable and realistic than rating scales or directly elicited willingness-to-pay (WTP)-type questions (James & Burton, 2003). We expect that BWS is likely to be more predictive of actual marketplace choices (Goodman, et al., 2006) as choice scenarios assist to identify the attributes that are in

the uppermost minds of respondents (Baek, Ham, & Yang, 2006). For more in-depth information about BWS in general please refer to Auger et al. (2007); Goodman, Lockshin & Cohen (2005); Hoek, Wong, Gendall, Louviere & Cong (2010); Jaeger, Jorgensen, Aaslyng & Bredie (2008) as good starting points.

To illustrate the flexibility in the BWS method selected noteworthy social science studies and their overall results are highlighted in table 1 below.

Table 1: Illustrative BWS Studies

Author	Title	BW Case	Overall results
Auger, et al. (2007)	Using Best: Worst Scaling Methodology to Investigate Consumer **Ethical Beliefs** across Countries	1	BWS is used to examine differences in attitudes of consumers towards social and ethical issues in six different countries with a total of more than 600 respondents. The results show that although there are differences, the most interesting results are the similarities. The most important issues were human rights, child labour and safe and good working conditions. Hence, some universal beliefs about social issues exist.
D'Alessandro & Winzar (2010)	Do students know best when it comes to assessment? A best/worst analysis of **assessment choices**	1	Language of schooling combined with work commitments to some extent determines a preference for more group work. Other students (local in particular) do not like to do group work or are indifferent. The paper highlights the difficulty of providing a homogenous education offering to a heterogeneous student population.
Louviere & Flynn (2010)	Using best-worst scaling choice experiments to measure public perceptions and preferences for **healthcare reform** in Australia	1	The BWS task forced respondents to discriminate between the 15 healthcare reform principles on offer. Quality and safety was the most important principle. It is suggested that researchers within the area of healthcare should consider applying the BW case 1 method in future studies.
Mueller, Lockshin, Saltman, & Blanford (2010)	Message on a bottle: The relative influence of wine back label information on **wine choice**	3	The importance of wine back label information relative to price was examined in a BWS study. Winery history, taste descriptions and food pairing were found to be of most importance to consumers. Ingredient information, on the other hand, had significant negative impact on one segment in particular.

3.4 BW Studies undertaken

In our work we have primarily focused on case 3 studies. Based on feedback from a 2006 non-pictorial, paper-based BWS study (Adamsen, 2006) where respondents expressed frustration at the repetitive nature of the task (15 choice scenarios with 4 options in each), we have generally applied a novel pictorial representation. The visual nature was chosen to improve task clarity (Adamsen, 2006; Paull, 2006) and to reduce the reliance on words (Louviere, Eagle, & Cohen, 2005, pp. 35-36). Pictorial representation also incorporates the fact that some attributes are difficult to verbally describe (Walley, et al., 1999). We started out with three different studies examining preferences for organic food products, and have subsequently completed another five online pictorial BWS studies for different baby skincare products, nappies and hand sanitiser.

Table 2: BWS studies undertaken (2009-2011)

#	Industry	Dates survey active	# of attri-butes	# of pro files	# of choice tasks	Pictorial represen-tation	Response rate	Data source	Sample
1	**Organic food (apples)**	**12/09-01/10**	**3**	**8**	**14**	**Yes**	**29 %**	**Griffith list**	**+18's***
2	Organic food (beef)	12/09-01/10	2	6	10	Yes	36 %	Griffith list	+18's*
3	Organic food (milk)	12/09-01/10	3	8	14	Yes	28 %	Griffith list	+18's*
4	Baby shampoo	02/11-03/11	4	9	12	Yes	36 %	Griffith list	(expectant) mothers*
5	Baby body wash	03/11	4	9	12	Yes	10 %	First Direct Solutions	(expectant) mothers*
6	Nappies	03/11	5	12	12	Yes	13 %	First Direct Solutions	(expectant) mothers*
7	Hand sanitiser	03/11	2	9	12	Yes	24 %	Griffith list	+18's*
8	Baby shampoo (different brands)	04/11	4	9	12	Yes	14 %	First Direct Solutions	(expectant) mothers*

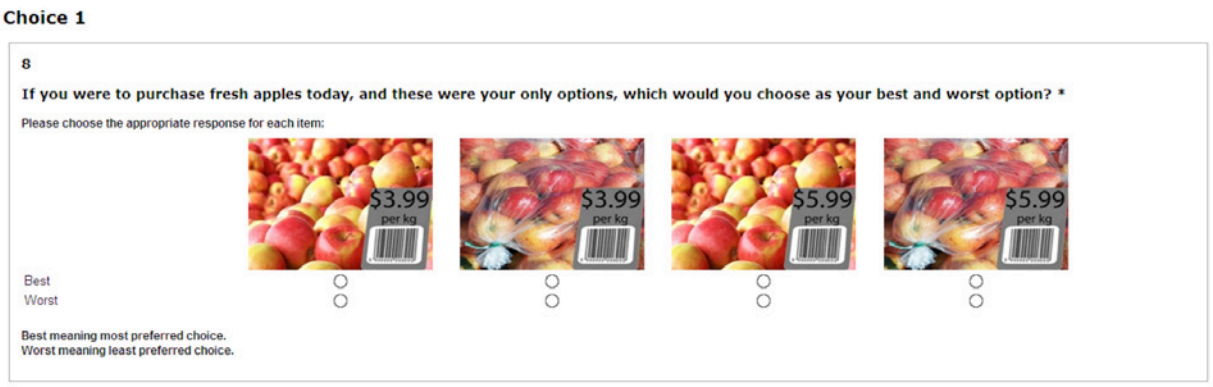
*All samples were largely representative of the Australian population when compared with ABS data in terms of family structure, income, level of education and all samples were Australia-wide.

Overall, the pictorial format was well received by the respondents and the response rates were higher than anticipated (based on an email list sampling). This suggests pictorial representation of BW studies is accepted by respondents. From here onwards the focus will be on the Apple study, as an example to illustrate the analysis and nuances of interpretation of both case 1 and case 3 BWS data. An example of the pictorial format can be seen in Figure 2 below.

4. ANALYSIS

BWS data can be analysed in different ways and the analytical approaches to be employed are highly dependent on the type of BWS (i.e. case 1, 2 or 3). Generally, for a case 1 BWS study a simple best minus worst calculation and associated investigation of the variance-covariance matrix to reveal consumer heterogeneity is often applied (Flynn, 2010; Steven Goodman, Lockshin, & Cohen, 2005; Mueller & Rungie, 2009), whereas type 3 BWS data is generally analysed through multinomial logit (MNL) and subsequently some clustering technique (e.g. latent class analysis (LCA)) to provide richer preference data. Below, in Figure 2, is an example of a case 3 BWS study examining preferences for (organic) apples. Profiles were developed pictorially with attribute and levels combinations based on a balanced incomplete block design (BIBD). In terms of analysis we have treated the data as a case 1 BWS study using B-W scores (i.e. by analysing the multiattribute profile into a single object without decomposed attributes) followed by a case 3 applying MNL analysis. Refer to Flynn (2010) for a much more detailed discussion of BWS data analysis approaches for case 1 and case 3 BWS studies.

Figure 2: Example of BW choice task (apple preferences)



4.1 Profile importance – the easy option providing an instant overview

For case 1 BWS data, the standard practice of calculating best minus worst (B-W) scores for each attribute or profile can be applied. For case 1 BWS, this approach has been found to provide results at the profile level comparable to the outcome of the more complicated MNL analysis of DCE data (Auger, et al., 2007; Flynn, et al., 2007; Hein, Jaeger, Carr, & Delahunty, 2008; Jaeger, et al., 2008; Lee, Soutar, & Louviere, 2007; Louviere & Islam, 2008). The B-W score then allows for ranking of the individual profiles. Positive values of B-W indicate that the given profile was chosen more frequently as best than worst and negative values reveal that the profile was chosen more frequently as worst. The average B-W scores are calculated by dividing the B-W score by the number of respondents and the frequency that each profile appears in the design of the choice set (r).

Another way to compare the profile importance is to derive ratio scores, by taking the square root after dividing the total B scores by the total W scores. The resulting coefficient measures the choice probability compared to the most important item (Auger, et al., 2007; Cohen, 2009; Flynn, et al., 2007; Lee, Soutar, & Louviere, 2008; Marley & Louviere, 2005). The square root of (B/W) for all profiles (sqrt(B/W)) are scaled by a factor, such that the most important profile with the highest square root (B/W) becomes Index 100. This allows for easy interpretation and comparison across profiles.

Case 3 BWS data, as we have for the apple study, can be treated as whole profiles and analysed as case 1 using the B-W score (refer to Table 3 below). The complexity of this analysis is low, and no specialist software package is required to undertake case 1 Best Worst analysis.

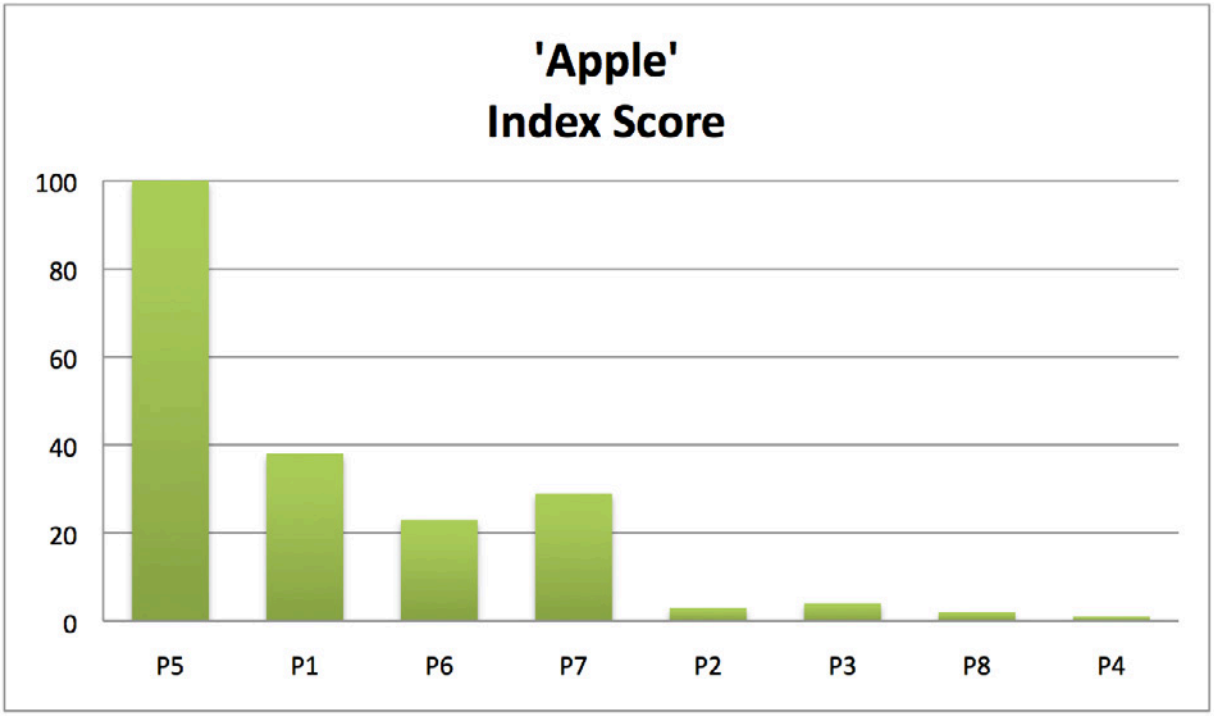
Table 3: Example of BW results in table form (preferences for organic apples treated as a case 1 study)

Profile #	Prod. method	Packaging	Price per kg AU$	Total Best	Total Worst	B-W score	Ave. B-W	Data source	Sample
5	B-W	Ratio	$3.99	1,499	12	**1,487**	6.38	11.18	**100**
1	Score	Index	$3.99	682	37	**645**	2.77	4.29	**38**
6	Organic	Bag	$3.99	489	74	**415**	1.78	2.57	**23**
7	Organic	Loose	$5.99	441	41	**400**	1.72	3.28	**29**
2	Conventional	Bag	$3.99	40	442	**-402**	-1.73	0.30	**3**
3	Conventional	Loose	$5.99	77	483	**-406**	-1.74	0.40	**4**
8	Organic	Bag	$5.99	27	698	**-671**	-2.88	0.20	**2**
4	Conventional	Bag	$5.99	7	1,475	**-1,468**	-6.30	0.07	**1**

Results in Table 3 indicate that loose, organic apples at a price of AU$3.99/kg was considered to be the best alternative by respondents. According to B-W analysis four apple alternatives received positive scores (all but one priced at AU$3.99/kg) while four apple alternatives were more frequently chosen as worst than best, subsequently scoring negatively. Case 1 BWS provides producers, retailers and marketers with aggregate level data providing information on alternatives that are most acceptable to the market and alternatives that would likely be rejected by the market. Refer to Figure 3 (representing the indexed B-W score from our case 3 study on apples; data shown in Table 3), where four alternatives receive positive B-W scores and four alternatives negative scores.

B-W scores can also be indexed, as sqrt(B/W), with the most preferred profile indexed at 100. This is visually depicted in Figure 3 below, where Profile 5 has an index of 100, and the second most preferred Profile, 1, is indexed at approximately 40. Profile 1, 6 and 7 are somewhat similar to consumers in terms of preference, whereas profiles 2, 3 and 8 are less preferred, evidenced through negative B-W scores (Table 3). Profile 4 ('conventional', AU$5.99 per kg, 'bag') is the least favoured profile, with an indexed score of 1, suggesting this profile has almost zero chance of being chosen.

Figure 3: Example of index scores (apple preferences)



4.2 Choice heterogeneity

Initial case 1 Best Worst analysis does not show any heterogeneity that may be present in the data. Hence calculations of variance and standard deviations can be used to further inform if choices have been consistent across all respondents, that is homogenous, or not. The standard deviations are calculated based on the individual B-W scores and the results are shown in Table 4 below, along with the a column indicating the ratio of standard deviation to the mean, which represents the extent of heterogeneity (high absolute ratios suggest greater heterogeneity).

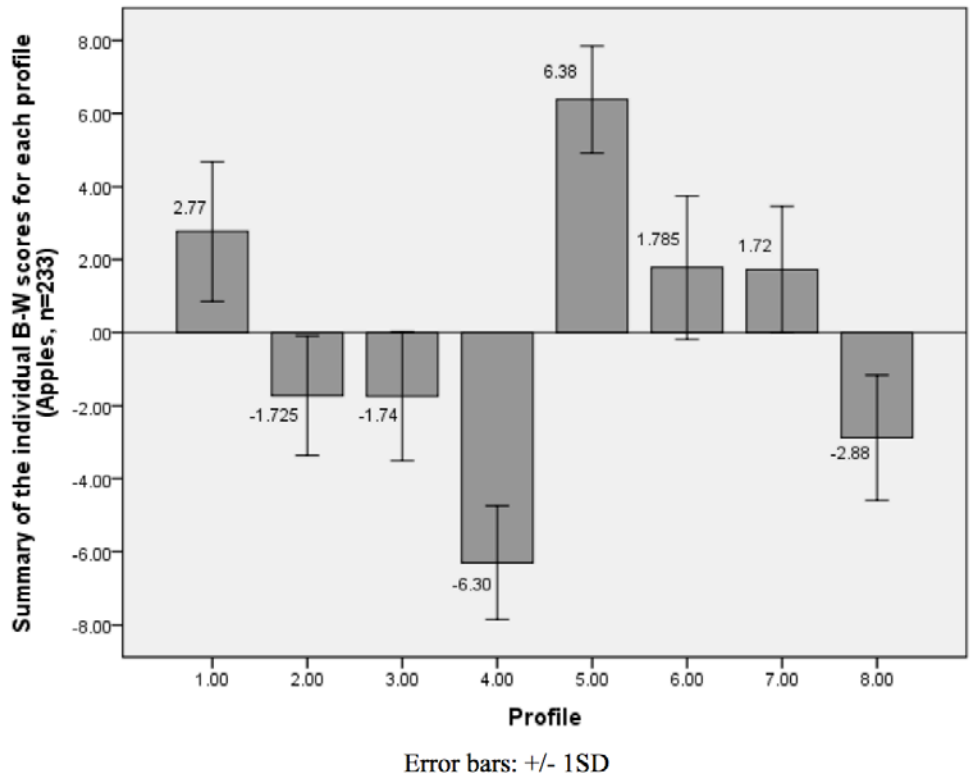
Table 4: Variance and standard deviation of profile importance

Profile	Attributes	Mean B-W	Var B-W	Stdev B-W	Stdev/ Mean
P6	Organic, bag, AU$3.99	1.78	3.86	1.97	1.11
P3	Conventional, loose, AU$5.99	-1.74	3.14	1.77	-1.02
P7	Organic, loose, AU$5.99	1.72	3.03	1.74	1.01
P2	Conventional, bag, AU$3.99	-1.73	2.67	1.64	-0.95
P1	Conventional, loose, AU$3.99	2.77	3.65	1.91	0.69
P8	Organic, bag, AU$5.99	-2.88	2.95	1.72	-0.60
P4	Conventional, bag, AU$5.99	-6.30	2.44	1.56	-0.25
P5	Organic, loose, AU$3.99	6.38	2.16	1.47	0.23

All profiles have standard deviations above one, as per the Table 4 above which signifies the existence of consumer heterogeneity for all profiles (Mueller & Rungie, 2009, p. 29). Some profiles tend towards homogeneity. These include the overall most preferred Profile 5 (Stdev/mean=0.23) as well as the overall least preferred Profile 4 (Stdev/mean=-0,25). Profiles 8 and 1 are also relatively homogeneous (Stdev/mean of -0.60 and 0.69 respectively). Other profiles, such as P6, P3 and P7 all have Stdev/mean > 1.0 indicating substantial disagreement, or heterogeneity, between respondents on the relative importance of apple profiles.

This information can also be presented graphically (see Figure 4), where the bars represent the average B-W scores, and the error bars show the standard deviation around the mean respectively. Hence the error bars span over two standard deviations. For better understanding the mean B-W score is also shown as a label next to the bar. Please note that

Figure 4: Profile importance and standard deviations

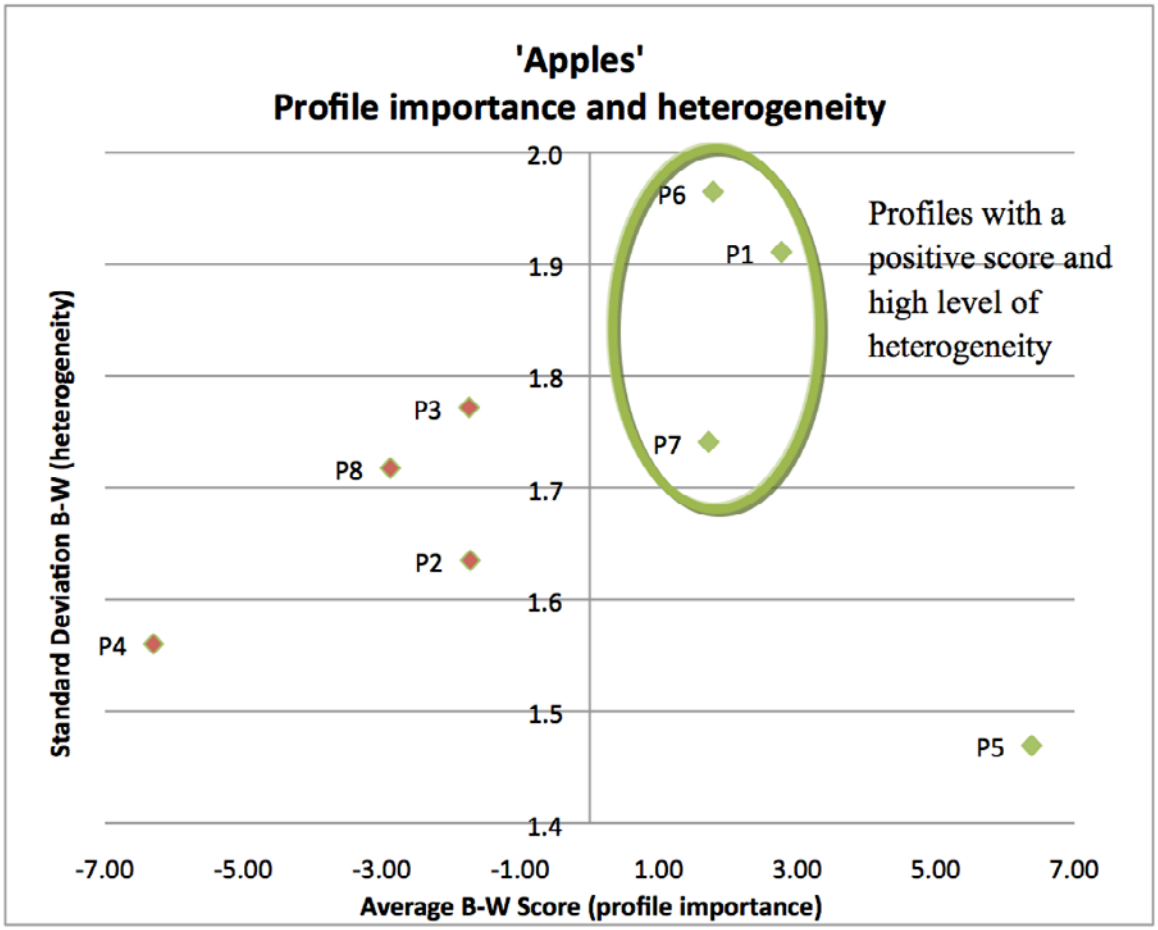


profiles are presented in numeric order in this output, starting from Profile 1 and ending with Profile 8. The ends of the error bars are marked with a line to increase visibility. Since each profile appears seven times in the design, the maximum value a profile can obtain is bound by +7 and the corresponding minimum value is -7. If a profile was chosen more often as best than worst, it will have a positive score, hence be above zero on the vertical axis (∑(B-W)>0), and similarly if chosen more often as worst it will be below zero on the axis. In this case Profiles 1, 5, 6 and 7 have positive scores, whereas Profiles 2, 3, 4 and 8 display negative scores (∑(B-W)<0). The size of the error bars relative to the size of the profile bars graphically represent the profile importance heterogeneity, and profiles with higher standard deviations, implying more heterogeneity, have longer whiskers.

The most important profile, Profile 5 (B-W score of 6.38), lies less than one standard deviation away from the highest possible score of +7. This indicates that a significant proportion of respondents always chose this profile as best when the profile appeared in the choice set. Contrary to this is Profile 4, the least preferred profile with a B-W score of -6.30. This profile's score is similarly less than one standard deviation away from the lowest possible score of -7, again indicating a high degree of consensus amongst respondents that this is the 'worst' choice in the majority of sets it was represented in.

Figure 5 below combines both the dimension of profile importance (B-W score) and heterogeneity (Stdev) in a standard scatter plot. Profiles which show homogeneity and high profile importance are the most favoured by the majority of consumers in the market. Profiles which have a positive mean B-W score and also show a high level of heterogeneity imply some respondent disagreement; hence the profiles may be important to a subset(s) of consumers. Those profiles are found in the upper, right corner of the graph (visualised by the green oval). In this case Profile 6 (organic, bag, AU$3.99), Profile 1 (conventional, loose, AU$3.99) and to a lesser extent Profile 7 (organic, loose, AU$5.99) may be preferred by a subset of consumers. Profiles like 3 (conventional, loose, AU$5.99) and 8 (organic, bag, AU$5.99) display relatively high Stdev's despite low B-W scores, and may be suitable to a smaller niche segment of consumers.

Figure 5: Positive B-W scores and high level of heterogeneity for apple profiles



4.3 Multinomial Logit – attribute importance

The overall B-W analyses presented so far have focused on average preferences and associated heterogeneity for *entire profiles* (based on data from a BWS type 3 experiment). This is the analytical procedure used for case 1 BWS studies and a discussion of the analysis of case 1 designs would end here. However, the advantage of case 3 BWS over case 1 is that it can be analysed using MNL and related cluster approaches such as latent class analysis to give separate importance weights for each attribute level, allowing the researcher to rank explode the number of observations to provide richer preference data. In a case 3 situation this analysis allows a direct assessment of the role each attribute level plays on stated choice preferences. A MNL analysis can help map the strength of each attribute and furthermore serve as a basis to estimate WTP for attribute levels, such as organic or loose apples, thereby assisting price setting decisions based on the attributes tested. In this particular example only the 'best' choices selected by respondents' form the basis of the analysis, therefore the data is treated as a standard DCE analysis. Other studies have also followed this approach (Bednarz, 2006), especially since the underlying psychological choice processes underlying best and worst choices may indeed differ.

The DCE has been widely applied in the area of food and drink marketing (see for example Lockshin, Jarvis, d'Hauteville, & Perrouty, 2006; Remaud, et al., 2008; Teratanavat & Hooker, 2006) and prior research has indicated that consumer choices made in a DCE format with visual shelf simulations are strongly related to actual market data (Mueller, Osidacz, et al., 2010). The advantage of using a DCE approach compared to other stated preference techniques is that it is possible to elicit preferences for the product attributes (packaging, price, organic versus conventional) rather than for the mere product as a whole (Bech & Gyrd-Hansen, 2005, p. 1079). A thorough description of the theory and statistics behind DCE's and Multinomial Logit (MNL) analysis can be found in Hensher, Rose & Greene (2005).

In regards to the specifics of the analysis employed to illustrate the apple data, NLogitsoftware (Econometric Software Australia, 2009) was used to estimate a Multinomial Logit (MNL) model as a linear function of the main attribute level effects. The price attribute was coded continuously, all other attributes were effects coded as per the general recommendation in the literature (Bech & Gyrd-Hansen, 2005; Lancsar & Louviere, 2008). Effects coding is argued to be advantageous over the more commonly used dummy coding approaches as the effects of all levels can be estimated uncorrelated with the intercept for the model (Bech & Gyrd-Hansen, 2005, p. 1082). When interpreting the results based on effects-coded variables, it is worth noting that the estimate of the referent level for the effects-coded attribute is the negative sum of the coefficients estimated for the other levels of that attribute (Bech & Gyrd-Hansen, 2005, p. 1080; Lancsar & Louviere, 2008, p. 670).

Table 5: Multinomial Logit model for the 'Apple' study

Study	Attribute	Đ-coefficient	WTP
Apple	Organic	**0.848*** (0.291)	1.244
	Loose	**1.392*** (0.039)	2.043
	Price (AU$/kg)	**-1.363*** (0.037)	

Standard errors are presented in (brackets) and * indicates the coefficients were significant at the p<0.01 level.
BIC = 1.189; Adjusted Pseudo R2 = 0.330
WTP for organic calculated as (0.848-(-0.848))/1.363=1.244 and similarly for loose as (1.392-(-1.392))/1.363=2.043).

Table 5 presents the results of a MNL model of the best choices for the apple data. The MNL model provides fixed Đ-coefficient estimates (referred to as part-worths, marginal utilities, or preference weights) for each of the attribute levels included in the model. The relative size and significance of these preference weights indicate the relative importance of the attribute level for alternative choice. In the case of the 'Apple'

study being able to purchase loose apples rather than apples in a bag (2.784[1]) rated higher than organic rather than conventional (1.696). Similarly, price is very important when consumers shop for apples, as every AU$1 increase in price per kilogram results in a 1.363 reduction in overall utility for the offering. This effect suggests that averaged across all consumers a kilogram of organic rather than conventional apples attracts an AU$1.244 (1.696/1.363) increase in price (shown in the Table as WTP in the last column). However, if the apples are both organic and offered in loose-weight, an increase in price of AU$3.29 ((1.696+2.784)/1.363) could theoretically be sustained without a change in overall utility for consumers.

4.4 Segmentation/clustering

The MNL model indicates the relative importance of different attributes for preferences for an "average" respondent across the sample, and the trade-offs respondents are willing to make between different attributes. In the apple example, the MNL model identified that all attributes included in the current study significantly impacted preferences with organic being valued as one of the least important attributes from a consumer point of view when compared to price and packaging. MNL models do not investigate the consumer heterogeneity this is present in the data. To identify preference heterogeneity, segmentation analysis can be applied based on the Best, Worst or Best-Worst choices obtained in case 3 BW studies. Examples of the application of Ward's hierarchical cluster method can be seen in Auger et al. (2007), whereas latent clustering has been utilised by Mueller and Rungie (2009). Clustering enables marketers to develop segments of customers based on choice preferences.

A latent class (LC) choice analysis extends the MNL approach to also consider how the relative importance of attributes varies in the respondent group by estimating detailed class memberships and part-worth utility parameters for different preference classes. In a LC analysis, the individual-level choices made (in our example in terms of the 'best' profile in every choice set) are regressed against the attribute levels that were presented in that choice set, along with socio-demographic variables (Remaud, et al., 2008, p. 8). The decision in regards to the number of classes is based on minimising the Bayesian Information Criterion (BIC) (Flynn, Louviere, Peters, et al., 2010).

In the illustrative analysis of our apple data, a three class solution was utilised, and six socio-demographic variables (gender, age, children, education, income and spending; all effects coded) were added one by one to the three-class solution to test for increased explanatory value. However, none of the added socio-demographic information improved the overall fit (i.e. the BIC was equivalent to that of the three class model without socio-demographic variables or higher) and the coefficient for the socio-demographic variables was not significant in predicting class membership ($p>0.05$). Hence the socio-demographic variables did not provide more explanatory power than the standard choice experiment attributes for this model. The estimates for the apple data for a LC model with three classes are presented in Table 6 below.

Table 6: Estimates of 'Apple' LC model

Class description	1. Price driven	2. Loose packaging preferred	3. Organic preferred
Class size	84.4%	7.1%	8.4%
Organic	1.237* (0.045)	-0.636* (0.069)	0.835* (0.062)
Loose	1.668* (0.054)	1.952* (0.109)	1.164* (0.057)
Price	-2.032* (0.063)	-1.325* (0.079)	0.348* (0.046)

Standard errors are presented in (brackets) and * indicates the coefficients were significant at the $p<0.01$ level.

[1] This is calculated as the difference between loose and bagged (due to effects coding, the coefficient for bagged is equal to -1 times the coefficient for loose, hence the difference between the two coefficients is equal to 1.392-(-1.392) = 2.784).

The overall fit of the model, adjusted pseudo-R2, was 0.66, representing a considerable improvement from the overall MNL model provided earlier (adjusted Pseudo-R2=0.33; Log Likelihood ratio test $p<0.01$). Overall, the most consistent feature of the LC model is being able to choose loose apples, as this attribute had significant and positive coefficients across all three classes. Nevertheless, it can be identified that the three classes differ in a number of ways.

First and foremost is a price-driven Class 1, where respondents prefer to choose their own loose apples (1.668), and they also like them to be produced organically (1.237), and are willing to pay a price premium (AU$1.22 per kg for organic apples; AU$2.86 per kg for organic, loose apples). This is the largest Class representing around 84% of the respondents.

The middle-ground is reflected in Class 2, where organic is seen as a negative product feature (-0.636) as is price (-1.325). This shows that this Class of consumers in fact prefer conventional apples and will not tolerate any price premium for organic products. However, being able to choose their apples is what matters the most to this group of respondents. Roughly 7% of respondents belong to this Class.

Last a marginal, and unexpected, positive price coefficient was estimated in Class 3. Taken together with positive coefficients for organic and loose also, we have discovered a Class of consumers who assign more importance and weight to first and foremost loose apples (1.164), and they also respond positively to organic apple offerings (0.835). This Class constitutes approximately 8% of the overall sample. However, it is worth noting that a positive price coefficient may indicate that the small proportion of respondents in this class either interpreted price as representing quality or did not clearly understand the task, hence caution should be used when describing this particular segment.

Section 4 has illustrated how case 1 and case 3 BWS studies can be analysed to gain insights into consumer preferences. A study examining stated choice preferences for certified organic apples was used to illustrate how whole of profile (or case 1 object) preferences can be examined, both at an aggregate level and to explore the nuances of heterogeneity between consumers. Next case 3 analysis was outlined to show how detailed preference data can be modelled to understand the individual attributes, estimate WTP and to examine segment preferences, all of which are critical in the new product development process. To conclude this paper we will now reflect on what we have learned from conducting a total of eight BWS studies over a two year period.

5. THE BEST, THE WORST AND THE FUTURE

In our first attempts employing visual representations, the organic food studies, we sought respondent feedback to gauge reactions to online surveys employing pictorial BWS. At the end of the first three pictorial online surveys (apples, beef and milk) respondents were asked in the very last question of the survey to provide feedback about the survey itself. We asked respondents to make any comments on the research and provided a box for them to enter comments indicating that any feedback would be appreciated. A total of 64 or 8.5% of all survey respondents chose to provide feedback on the online surveys. In general, the study received positive feedback from respondents. Overall 17% of comments were negative and the remaining 83% of comments were positive.

Many respondents commented on the ease of the online survey employing BWS with some respondents stating the survey was easy to complete, quick, interesting, and user-friendly. This is evident in the following respondent comment *"Nice approach. I think this is the first time I've seen this kind of imagery in a web survey."* While other respondents commented *"A fast survey is a good survey", "More interesting doing a survey that has pictures"* and *"Great graphics made it easy..."*.

Some respondents commented on the costs involved in downloading the data (as pictorial surveys take up more virtual space and associated download costs), an issue the research team had not considered during the survey design process. Consider the following respondent comment *"despite the presence of images, the survey was not too costly to my mobile broadband data charges. Perhaps some indication at the start of the survey about expected megabytes of data required to complete would be helpful for those on expensive connections, especially regional Australia where only Telstra may be available. You may get a higher participation rate by providing this information (without guarantees of course!)"*. A further issue not considered by the research team in the survey design stage was that some respondents may elect to complete a survey using a mobile data device (e.g. iPhone). One respondent commented *"Slightly difficult to choose answers on an iPhone"*. These comments suggest that market researchers seeking to maximise response rates must consider all survey platforms and that if maximum response is desired alternative survey versions along with data cost information (e.g. number of megabytes to download) should be provided for respondents.

As noted earlier not all respondents were generally positive towards the online survey. Negative comments focussed on the repetitive nature of the BWS choice tasks (pictorial or not). Consider the following respondent comment *"The number of questions were toooo many. Same thing over and over, the temptation is to randomly pick, just so to end it. If it was not for the rewards, I would not have finished it"* while another respondent commented *"the scenarios seemed very similar..."*. Moreover, pilot testing on one baby shampoo survey involving 20 choice scenarios with 4 product alternatives per scenario was deemed too long and repetitive and received extreme negativity forcing a re-design resulting in 12 choice scenarios. Subsequent testing of this shorter design (12 choices with 3 alternatives for each) received positive feedback. We consequently aimed to limit the number of choice tasks in future studies, and overall respondent feedback certainly indicates that 10 to 14 choice tasks in one online survey should be considered as a limit to avoid respondent fatigue and to limit drop-out rates.

From a researchers' perspective case 3 BWS is a powerful method yielding data that can be used to predict optimum attribute levels (e.g. prices that people are willing to pay). Data needs to be organised manually prior to undertaking data analysis, however this process is relatively quick. Aggregate data reads, showing the most and least preferred profiles, can be obtained in one working day following closure of an online study. In contrast to techniques such as DCE, the advantage of BWS is the amount of information gathered; since the design ensures more than one paired comparisons is obtained from each choice set. Difficulties faced by researchers employing the BWS method centre on research design and the lack of textbooks explaining this technique. It is essential that researchers understand the key attributes for a product and that sufficient attribute levels are selected. For example, omission of a key attribute or a price point that is reflected in the market place would jeopardise the results obtained in the BWS method. However, at the same time researchers are limited in regards to the number of attributes, and related levels, under investigation, as the choice scenarios quickly become too large, complex and time-consuming for respondents. Hence it is a balancing act between what information is essential to address the research question and the necessity to find a pragmatic design to match this.

Overall, we do believe the BWS method has a significant potential to improve predictability in market research – the response rate and positive feedback from participants is encouraging!

6.1 Limitations and future research

BWS studies are bound by the attributes studied and our experience suggests the number of attributes that can be studied must be kept in check to limit the number of choice scenarios. Most case 1 BWS studies in marketing are based on Balanced Incomplete Block (BIB) designs and most BIBs are for just two levels of each attribute (Zikmund, et al., 2011, p. 535). There are some three-level BIB designs published, but these are sparse. The reality then is that the researcher must choose between reducing the number of levels in the attribute(s) or simply accept a design that is much larger than necessary.

A key limitation of the BWS method is that researchers are not able to comment beyond the attributes studied. In the case of the organic apple study we are not able to comment on the appearance (cosmetic quality) of organic apples compared with conventionally produced apples. Our study findings are limited to price, packaging and production method. In the case of the baby shampoo survey our results are restricted to three brands despite our knowledge that the baby shampoo market consists of more than three brands. Presenting participants with a larger number of brands would have resulted in too large an experimental design for the available sample.

Though BWS has been shown to produce results that are closer to actual in-market behaviour than standard rating techniques (Auger, Devinney, & Louviere, 2007; Goodman, Lockshin, & Cohen, 2006), future research employing mixed methods should empirically test to what extent BWS minimises biases such as SDR, acquiescence, scalar equivalence, ERS and IS responses when investigating consumer behaviour.

Market researchers focusing on multi attribute studies are advised to employ pictorial BW choice tasks as depicted in this paper in Figure 2 as respondents report difficulties in processing information on multiple attributes in word form. A pictorial representation such as that depicted in Figure 2 allows respondents to easily process information on multiple attributes simultaneously much as they would in a supermarket shelf setting. However, it should be noted that pictorial representation may also induce bias because of respondent variation in interpretation and/or perceptions of the images used, and the colour scheme may also influence results. Nevertheless, we do recommend using this visual layout when practically possible, due to the positive feedback from respondents. As one respondent commented *"A survey with pictures is better than a survey with 1000 words"*.

REFERENCES

Adamsen, J. M. (2006). *To be or not to be an organic consumer: A cross-cultural and methodological comparison. The case of Australia and Denmark*. Unpublished Thesis. Griffith University.

Alriksson, S., & Öberg, T. (2008). Conjoint analysis for environmental evaluation. *Environmental Science and Pollution Research, 15*(3), 244-257.

American Marketing Association. (1992). *Conjoint analysis: A guide for designing and interpreting conjoint studies*: AMA.

Auger, P., Devinney, T. M., & Louviere, J. J. (2007). Using Best-Worst scaling methodology to investigate consumer ethical beliefs across countries. *Journal of Business Ethics, 70*(3), 299-326.

Baek, S. H., Ham, S., & Yang, I. S. (2006). A cross-cultural comparison of fast food restaurant selection criteria between Korean and Filipino college students. *International Journal of Hospitality Management, 25*, 683-698.

Barkworth, L., Hibbert, S., Horne, S., & Tagg, S. (2002). Giving at Risk? Examining perceived risk and blood donation behaviour. *Journal of Marketing Management, 18*(9), 905 - 922.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2005). Response biases in marketing research. *The Handbook of Market Research: Do's and Dont's* (pp. 204-237): Sage Publications.

Bednarz, A. (2006). *Best-Worst scaling and its relationship with multinomial logit*. University of South Australia.

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of context and style: two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76*, 186-204.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(4), 608 - 628.

Biological Farmers of Australia. (2010). *Australian Organic Market Report 2010*.

Cattin, P., & Wittink, D. R. (1982). Commercial use of conjoint analysis: A survey. *The Journal of Marketing, 46*(3), 44-53.

Chen, C., Lee, S.-y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among east Asian and North American students. *Psychological Science, 6*(3), 170-175.

Chrzan, K., & Skrapits, M. (1996). *Best/worst conjoint analysis: An empirical comparison with a full profile choice-based conjoint experiment*. Paper presented at the INFORMS Marketing Science Conference.

Cohen, E. (2009). Applying best-worst scaling to wine marketing. *International Journal of Wine Business Research, 21*(1), 8-23.

Cohen, S., & Markowitz, P. (2002). *Renewing market segmentation: Some new tools to correct old problems*. Paper presented at the ESOMAR 2002 Amsterdam, The Netherlands.

Cohen, S., & Neira, L. (2003). *Measuring preferences for product benefits across countries: Overcoming scale usage bias with maximum difference scaling*. Paper presented at the Latin American Conference of ESOMAR, Punta del Este, Uruguay.

D'Alessandro, S., & Winzar, H. (2010). *Do students know best when it comes to assessment? A best/worst analysis of assessment choices*. Paper presented at the ANZMAC, Christchurch.

Finn, A., & Louviere, J. J. (1992). Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy & Marketing, 11*(2), 12-25.

Flynn, T. N. (2010). Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. *Expert Review of Pharmacoeconomics & Outcomes Research,* 10(3), 259-259-267.

Flynn, T. N., Louviere, J. 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*(1), 171-189.

Goodman, S. (2009). An international comparison of retail consumer wine choice. *International Journal of Wine Business Research, 21*(1), 41-49.

Goodman, S., Lockshin, L., & Cohen, E. (2006). *Using the Best-Worst method to examine market segments and identify different influences of consumer choice.* Paper presented at the 3rd International Wine Business and Marketing Research Conference.

Goodman, S., Lockshin, L., & Cohen, E. (2005). *A simple method to determine drinks and wine style preferences* Paper presented at the Second Annual International Wine Marketing Symposium.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *The Journal of Consumer Research, 5*(2), 103-123.

Hein, K. A., Jaeger, S. R., Carr, B. T., & Delahunty, C. M. (2008). Comparison of five common acceptance and preference methods. *Food Quality and Preference, 19*(7), 651-661.

Hensher, D. A. (2010). Hypothetical bias, choice experiments and willingness to pay. *Transportation Research Part B: Methodological, 44*(6), 735-752.

Hoek, J., Wong, C., Gendall, P., Louviere, J., & Cong, K. (2010). Effects of dissuasive packaging on young adult smokers. *Tobacco Control*(2010-10-21).

Holdershaw, J. L., Gendall, P. J., & Wright, M. J. (2011). Predicting blood donation behaviour: Further application of the theory of planned behaviour. *Journal of Social Marketing, 1*(2), 1-1.

Huber, J., Wittink, D. R., Johnson, R. M., & Miller, R. (1992). *Learning effects in preference tasks: Choice-based versus standard conjoint.* Sawtooth Software Inc.

Jaeger, S. R., Hedderley, D., & MacFie, H. J. H. (2001). Methodological issues in conjoint analysis: a case study. *European Journal of Marketing, 35*(11), 1217-1239.

Jaeger, S. R., Jorgensen, A. S., Aaslyng, M. D., & Bredie, W. L. P. (2008). Best-worst scaling: An introduction and initial comparison with monadic rating for preference elicitation with food products. *Food Quality and Preference, 19*(6), 579-588.

James, S., & Burton, M. (2003). Consumer preferences for GM food and other attributes of the food system. *Australian Journal of Agricultural and Resource Economics, 47*(4), 501-518.

Karniouchina, E. V., Moore, W. L., van der Rhee, B., & Verma, R. (2009). Issues in the use of ratings-based versus choice-based conjoint analysis in operations management research. *European Journal of Operational Research, 197*(1), 340-348.

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*(4), 335-347.

Lee, J. A., Soutar, G. N., & Louviere, J. (2007). Measuring values using best-worst scaling: The LOV example. *Psychology & Marketing, 24,* 1043-1058.

Lockie, S., Lyons, K., Lawrence, G., & Grice, J. (2004). Choosing organics: a path analysis of factors underlying the selection of organic food among Australian consumers. *Appetite, 43*(2), 135-146.

Loomis, J. (2011). What's to know about hypothetical bias in stated preference valuation studies? *Journal of Economic Surveys, 25*(2), 363-370.

Louviere, J. J., Eagle, T. C., & Cohen, S. H. (2005). *Conjoint analysis: Methods, myths and much more.* Unpublished Working Paper. Centre for the Study of Choice, Faculty of Business, University of Technology Sydney

Louviere, J. J., & Flynn, T. N. (2010). Using best-worst scaling choice experiments to measure public perceptions and preferences for healthcare reform in Australia.(Original Research Article). *The Patient: Patient-Centered Outcomes Research, 3*(4), 275(279).

Louviere, J. 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, 61(9), 903–911.*

Louviere, J. J., Swait, J., & Anderson, D. (1995). *Best/ Worst conjoint: A new preference elicitation method to simultaneously identify overall attribute importance and attribute level partworths.* University of Sydney.

Lusk, J. L., McLaughlin, L., & Jaeger, S. R. (2007). Strategy and response to purchase intention questions. *Marketing Letters, 18*(1-2), 31-44.

Magnusson, M. K., Arvola, A., Hursti, U.-K. K., Åberg, L., & Sjödén, P.-O. (2001). Attitudes towards organic foods among Swedish consumers *British Food Journal 103*(3), 209-227.

Marley, A. A. J., & Louviere, J. J. (2005). Some probalistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology, 49*(6), 464-480.

McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research, 20*(1), 60-103.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 252). New York: Academic Press.

Mick, D. G. (1996). Are studies of dark side variables confounded by socially desirable responding? The case of materialism. *Journal of Consumer Research, 23*(2), 106.

Moore, W. L. (2004). A cross-validity comparison of rating-based and choice-based conjoint analysis models. *International Journal of Research in Marketing, 21*(3), 299-312.

Mueller, S., Lockshin, L., Saltman, Y., & Blanford, J. (2010). Message on a bottle: The relative influence of wine back label information on wine choice. *Food Quality and Preference, 21*(1), 22-32.

Mueller, S., & Rungie, C. (2009). Is there more information in best-worst choice data?: Using the attitude heterogeneity structure to identify consumer segments. *International Journal of Wine Business Research, 21*(1), 24-40.

Murphy, J., Allen, P. G., Stevens, T., & Weatherhead, D. (2005). A Meta-analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental and Resource Economics, 30*(3), 313-325.

Netzer, O., Toubia, O., Bradlow, E., Dahan, E., Evgeniou, T., Feinberg, F., & Rao, V.R. (2008). Beyond conjoint analysis: Advances in preference measurement. *Marketing Letters, 19*(3), 337-354.

New York AMA Communication Services Inc. (2013). *GRIT report* New York: New York AMA Communication Services Inc.

Oliphant, K., Eagle, T. C., Louviere, J. J., & Anderson, D. (1992). *Cross-task comparison of ratings-based and choice-based conjoint.* Paper presented at the Sawtooth Software Conference, Ketchum.

Özdemir, S., Johnson, F. R., & Hauber, A. B. (2009). Hypothetical bias, cheap talk, and stated willingness to pay for health care. *Journal of Health Economics, 28*(4), 894-901.

Paull, J. (2006). *Provenance, purity & price premiums: Consumer valuations of organic & place-of-origin food labelling.* University of Tasmania.

Potoglou, D., Burge, P., Flynn, T., Netten, A., Malley, J., Forder, J., et al. (2011). Best-worst scaling vs. discrete choice experiments: An empirical comparison using social care data. *Social Science & Medicine, 72*(10), 1717-1727.

Rose, J. M., Hensher, D. A., Greene, W. H., & Washington, S. P. (2011). Attribute exclusion strategies in airline choice: accounting for exogenous information on decision maker processing strategies in models of discrete choice. *Transportmetrica*, 8 (5), 344-360.

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association, 96*(453), 20-31.

Rundle-Thiele, S. (2009). Bridging the gap between claimed and actual behaviour. The role of observational research. *Qualitative Market Research: An International Journal, 12*(3), 295-306.

Saris, W.E., Revilla, M., Krosnick, J.A., and Schaeffer, E. M. (2010). Comparing questions with agree/ disagree response options to questions with item-specific response options (2010). *Survey Research Methods, 4*(1), 61-79.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research *Journal of Consumer Research, 25*(1), 78-91.

Steenkamp, J.-B. E. M., Hofstede, F. t., & Wedel, M. (1999). A cross-national investigation into the individual and national cultural antecedents of consumer innovativeness. *The Journal of Marketing, 63*(2), 55-69.

Steenkamp, J.-B. E. M., & Ter Hofstede, F. (2002). International market segmentation: Issues and perspective. *International Journal of Research in Marketing, 19*(3), 185-213.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review, 34*(4), 273-314.

Vermeir, I., & Verbeke, W. (2006). Sustainable food consumption: Exploring the consumer "attitude – behavioral intention" gap. *Journal of Agricultural and Environmental Ethics, 19*(2), 169-194.

Walley, K., Parsons, S., & Bland, M. (1999). Quality assurance and the consumer: A conjoint study. *British Food Journal 101*(2), 148-162.

Watson, D. (1992). Correcting for acquiescent response bias in the absence of a balanced scale. *Sociological Methods & Research, 21*(1), 52-88.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236-247.

Welkenhuysen-Gybels, J., Billiet, J., & Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items *Journal Of Cross-cultural Psychology, 34*(6), 702-722.

Wittink, D., & Walsh, J. (1988). *Conjoint analysis: Its reliability, validity, and usefulness*. Paper presented at the Sawtooth Software Conference, Ketchum.

Zikmund, W., Ward, S., Lowe, B., & Winzar, H. (2007). *Marketing Research Asia Pacific Edition* (1 ed.). Melbourne: Thomson Learning.

Zikmund, W. G., Ward, S., Lowe, B., Winzar, H., & Babin, B. J. (2011). *Marketing Research* (2nd Asia-Pacific Edition ed.). Melbourne: Cengage Learning Australia Pty Limited

Zinkhan, G. M., & Carlson, L. (1995). Green advertising and the reluctant consumer. *Journal of Advertising, 24*(2), 1.

ACKNOWLEDGEMENTS

The authors would first and foremost like to thank Peter Vitartas for valuable input to this paper and to the blind reviewers who assisted on earlier versions of this manuscript. Furthermore, would we like to thank the AMSRS for the opportunity to present this work at the 2011 conference. We also need to thank Christina Mehnert for great assistance in the development of the online surveys and subsequent data collection.