

Managing groundwater in a mining region: an opportunity to compare best-worst and referendum data

Darla Hatton MacDonald†, John M. Rose, Robert J. Johnston, Rosalind H. Bark and Jodie Pritchard

† Corresponding author: Darla Hatton MacDonald (Darla.HattonMacDonald@utas.edu.au),
Tasmanian School of Business & Economics, University of Tasmania, Sandy Bay, Tasmania, Australia.

John M. Rose, Centre for Business Intelligence and Data Analytics, University of Technology Sydney,
Sydney, New South Wales, Australia.

Robert J. Johnston, Department of Economics, Clark University, Worcester, Massachusetts, USA.

Rosalind H. Bark, School of Environmental Sciences, University of East Anglia, Norwich, UK.

Jodie Pritchard, Biodiversity, Ecosystem Knowledge and Services Program, CSIRO Waite Campus,
Urrbrae, South Australia, Australia.

Abstract

In nonmarket valuation, practitioners must choose a format for the valuation questions. A common approach in discrete choice experiments is the ‘pick-one’ format, often with two alternative policy proposals and a status quo from which the respondent selects. Other proposed formats, include best-worst elicitation, where respondents are asked to indicate their most and least favoured alternative from a set. Although best-worst formats can offer efficiency in data collection, they can also lead to responses that are difficult to reconcile with neoclassical welfare estimation. The current article explores methodological issues surrounding the use of pick-one versus best-worst data for nonmarket valuation, focusing on framing and status quo effects that may occur within three-alternative discrete choice experiments. We illustrate these issues using a case study of surplus groundwater use from Western Australian mining. Results identify concerns that may render best-worst data unsuitable for welfare estimation, including a prevalence of serial choices in which the status quo is universally chosen as the worst alternative, rendering part of the choice process deterministic. Asymmetry of preferences and serial choices can be obscured when models are estimated using ‘naively’ pooled best-worst data. Results suggest that caution is warranted when using best-worst data for valuation, even when pooled results appear satisfactory.

Key words: Aboriginal cultural sites, best-worst scaling, groundwater, habitat, mining, water resources, willingness to pay.

1. Introduction

Stated preference methods are used widely to elicit public preferences and estimate willingness to pay (WTP) for changes in environmental quality or ecosystem services (henceforth, ‘environmental valuation’). Early stated preference applications such as contingent valuation (CV) relied on relatively simple elicitation mechanisms including open-ended, payment cards, dichotomous choice, and double-bounded questions (Smith 2006). CV is still widely applied, often employing the commonly prescribed dichotomous choice format (Boyle 2017). In the environmental valuation literature, discrete choice experiments (DCEs), also called choice models (Hanley et al. 2001), are increasingly prevalent. In a DCE survey, respondents are asked to consider a set of multi-attribute choice alternatives that alter a set of environmental and other conditions (typically including household cost), and to indicate the alternative they would choose or prefer. Among the methodological choices faced by those applying DCEs is response format, or the format of the question used to elicit

preference information from respondents (Johnston et al. 2017; Petrolia et al. 2018; Yangui et al. 2019).

The most common response format in environmental valuation DCEs is a 'pick-one' format framed as a referendum. This often involves two alternative policy proposals (A and B) and a status quo (Hensher et al. 2015). However, there are many other common response formats, including: multi-attribute dichotomous (binary) choice (Brefle and Rowe, 2002); and best-worst (BW) elicitation (Scarpa et al. 2011; Louviere et al. 2015). Among the central questions when choosing among these formats is whether model results (e.g. preference and welfare estimates) are robust (Giergiczny et al. 2014). If results are robust to alternative response formats, practitioners would ideally be free to choose the format best suited to their perceived research needs, considering issues such as the decision context and available sample sizes. For example, BW could be used with smaller sample sizes (Scarpa et al. 2011), without concern that the resulting welfare estimates might be sensitive to this choice. Past work has been mixed regarding the sensitivity of welfare and preference estimates to stated-preference response formats (see review in Petrolia et al. 2018).

This paper responds to a call by Johnston et al. (2017) for research that compares the performance of alternative response formats. Here, we focus explicitly on the performance of BW elicitation compared to the pick-one format commonly used in three-alternative valuation DCEs. The relevance of these analyses is underscored by inconsistent findings and recommendations in the literature related to the use of BW data for preference and welfare estimation. Although some work has raised cautions about the tendency of this format to generate inconsistent preference and error variance estimates across the spectrum of best and worst responses (e.g. Giergiczny et al. 2014; Rose 2013), other analyses have found that BW formats can provide reasonable preference information similar to that generated by alternative formats (e.g. Yangui et al. 2019; Petrolia et al. 2018). The juxtaposition of these seemingly inconsistent findings combined with a 'relatively young' literature in this area (Petrolia et al. 2018, p. 367), suggests the need for additional work to evaluate the validity of BW elicitation within valuation contexts such as the ubiquitous three-alternative DCE.

This analysis is conducted using a DCE on preferences of people living in Australia for uses of surplus groundwater from deep mining operations in Western Australia. Using split-sample BW and pick-one response data from an otherwise identical three-alternative DCEs, we compare models in terms of WTP estimates, error variances, and the consistency of best and worst responses. We give particular attention to framing and serial status quo effects that can be obscured within pooled BW data. The model is estimated in WTP space to ameliorate welfare-estimation challenges that can be encountered with preference-space models (Train and Weeks 2005; Scarpa et al. 2008).

The paper proceeds as follows. We first provide an overview of elicitation approaches to place our study in context of the broader literature. We then describe the case study, followed by a presentation of DCE design, data, and analytical methods for model estimation and hypotheses tests. Foreshadowing the results that follow, we find evidence of preference asymmetries and non-trading behaviour within BW data that jeopardise the validity of welfare estimation. Although some estimates support a preliminary finding that welfare estimates are similar across response formats, the validity of this statement is belied by inconsistencies between best and worst responses and poorly conditioned data from worst responses. Results suggest that caution is warranted when using BW data for valuation, even when pooled results appear satisfactory.

2. Literature review

If one assumes strict neoclassical decision-making with perfect and consistent information processing, then the format of DCEs should be largely irrelevant to the choice process and model outcomes, as long as the questions are consistent with the same underlying random utility model (Johnston and Swallow 1999; Swait and Adamowicz 2001; Meyerhoff et al. 2015). One of the few systematic comparisons of BW to other common elicitation formats in the environmental valuation literature is that of Petrolia et al. (2018), conducted in the context of ecosystem service valuation within the United States. Their analysis compares the results of one-shot single dichotomous choice, repeated pick-one DCE and BW case 3 applications and finds little difference in parameter and welfare estimates after allowing for differences in scale and the order of questions.

However, a broad and increasing range of evidence challenges whether the format of the DCE has an impact on the choice process. For example, information load and design dimensions of DCEs can have mixed influence on attribute processing strategies and results (DeShazo and Fermo 2002; Hensher 2006, b; Meyerhoff et al. 2015). Moreover, the properties of different types of stated preference elicitation formats can influence the extent to which questions encourage truthful preference revelation (Carson and Groves 2007; Vossler et al. 2012). These issues render the choice of elicitation format an important aspect of contemporary stated preference design (Johnston et al. 2017).

There are advantages and disadvantages to all common DCE response formats (Petrolia et al. 2018). For example, one-shot, binary referendum questions (i.e. two-alternative, pick-one formats) can have desirable properties (Carson and Groves 2007). Theoretical conditions for incentive compatibility in three-alternative DCEs are narrow and often unlikely to hold in practice (Collins and Vossler 2009; Vossler et al. 2012). However, one-shot binary questions provide less information per question/survey and may be subject to 'yea-saying', leading to inflated estimates of WTP (Hanley et al. 2001).

BW formats (or scaling), in contrast, seek to obtain a complete preference ordering over available choice alternatives through a set of questions asking respondents to indicate their best (most favoured) and worst (least favoured) alternative (Louviere et al. 2015). Although BW elicitation lacks the desirable incentive properties of appropriately designed one-shot binary questions, they provide more information per question than two-alternative or three-alternative pick-one formats. It has also been argued that the relatively low cognitive effort associated with many types of BW formats can enhance choice consistency and provide more accurate preference information compared to other approaches (e.g. complete ranking) that provide similar amounts of information (Flynn et al. 2007).

Various forms of BW elicitation have been applied in fields such as health care (Flynn et al. 2007), food choice (Lusk and Briggeman 2009; Yangui et al. 2019), transportation (Beck et al. 2017), and environmental economics (Soto et al. 2018; Petrolia et al. 2018). To clarify terminology, there are three primary types of BW elicitation, often denoted as case 1, 2, or 3 (Louviere et al. 2015). Case 3 corresponds closely to the traditional pick-one DCE format used in environmental valuation and is the format evaluated here. Within this format, attributes and choice alternatives are organised in a parallel fashion to those in a pick-one DCE, with respondents asked to consider two or more alternatives, each comprised of multiple attributes. However, instead of choosing a single preferred alternative, respondents in a case 3 BW question are asked to choose their most (best) and least (worst) favoured alternatives. Additional stages are required when there are more than three alternatives, in which case, the initially chosen best and worst alternatives are removed from the set, and respondents are then asked to identify the best and worst alternatives from the remaining set (Louviere et al. 2015). This process continues until a complete ranking is obtained. The resulting response data can be exploded into a set of pseudo-observations that predict choices over all

possible combinations of alternatives (Scarpa et al. 2011; Rose 2013). Operationally, this approach is attractive for efficiently gathering information with respect to the ranking of alternatives and the ease with which respondents select extremes (Marley and Louviere 2005).

Although applications of BW approaches tend to emphasise advantages, there are also potential disadvantages. For example, BW elicitation can lead to differences in preferences and error variances between best and worst choices, and across different iterations of BW choices (Rose 2013; Giergiczny et al. 2014). As described above, case 3 is often used to generate an exploded set of independent pseudo-observations, thereby enabling more robust parameter estimates for a given sample size (see, e.g. Vermeulen et al. 2011). However, multiple researchers have questioned such an approach, given the possibility that respondents may exhibit different preferences and/or error variances over the various pseudo-observations. For example, Collins and Rose (2011) and Scarpa et al. (2011) examine potential error variance differences between best and worst observations; whilst Rose (2013) and Giergiczny et al. (2014) examine error variance and preference differences between the two choices. At least some differences are found in all cases.

Explaining these findings, Rose (2013) argues that best and worst choices may reflect different response frames, one positive and one negative. As such, there is no reason why one would assume that the preferences (and error variances) obtained from one type of question should mirror that of the other, as answers to best and worst questions may arise from different data generating processes. That is, the framing of BW questions may lead respondents to use asymmetric preferences when evaluating positive (best) and negative (worst) dimensions (Rose 2013). This mirrors earlier findings in the CV literature (Johnston and Swallow 1999). Given the possibility of asymmetric preferences, Rose (2013) questions the common 'naïve' pooling of BW data to obtain a single vector of preference weights. Following a similar line of argument, Dyachenko et al. (2014) find that the data generation process differs between best and worst responses, depending on the response sequence.

There is also a possibility that certain types of response behaviours that are inconsistent with fully compensatory utility maximisation may be magnified (or diminished) under BW elicitation, compared to other response formats. These patterns might be influenced by framing effects of the type described above. For example, whilst some past work has evaluated status quo effects in BW data (Petrolia et al. 2018), we are not aware of any research that considers whether these effects exist and/or vary across best and worst responses. These concerns imply that additional evaluation within environmental valuation DCEs is required before BW elicitation can be recommended for broader use.

3. Case study

Coal and gas extraction, metal ore mining, and other types of mining are important sectors of the Australian economy in terms of export earnings (ABS 2015), with before-tax earnings amounting to A\$83.6 billion in 2016 (ABS 2018). Our study site is in the Pilbara, Western Australia, located in the north-western corner of the State (Figure 1). The region is sparsely populated with large cattle properties, mining operations (primarily iron ore), small resource-based towns, and Aboriginal communities.

This case study addresses groundwater management options related to mining operations. This management issue arises when excavations for metal ores intersect aquifers, requiring the removal of groundwater before mining can proceed. This is also a growing concern for coal seam gas extraction, see Currell (2015). The process of dewatering for mining operations in the Pilbara

requires 180 to 200 GL/year of groundwater water to be extracted. The disposal (or use) of this extracted groundwater can have multiple environmental impacts, depending on the disposal option (ABC News 2013). Approved groundwater disposal options in Western Australia include: (1) reuse of water in mining and processing either in situ or at nearby mines; (2) local groundwater recharge; (3) direct discharge to nearby wetlands, rivers, drains, or drainage lines; (4) use in irrigation; and (5) water storage (Government of Western Australia, 2000).

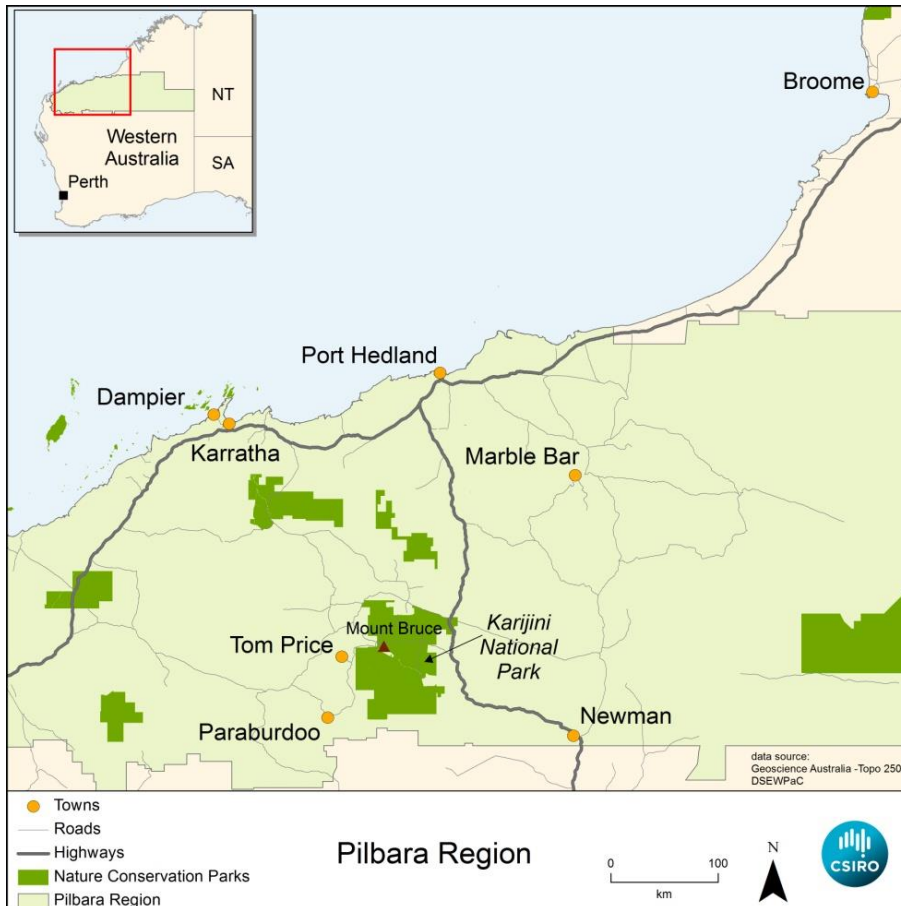


Figure 1 Map of the Pilbara region, Western Australia

New mine sites are subject to social, political, technical, and environmental constraints. In the Pilbara, large tracts of land are held by the Crown, in Native Title by Aboriginal communities, as well as by private landholders and private leaseholders in extensive cattle stations and mineral titles. At the State level, the Government of Western Australia’s Department of Water recognises that ‘mining projects can have significant impact on groundwater and surface water resources, and their associated values’ (WA DOW 2012). The Department’s management objectives include minimising the adverse effects of abstraction and release, optimising groundwater extraction, and accounting for the regional cumulative effects of water resources management from mining. Information on the values people hold for the various potential uses of surplus groundwater, and the outcomes of those uses, is necessary to understand the net social benefits of alternative management options. Hence, this DCE was designed to explore Australians’ preferences for different surplus groundwater use options, with parallel questionnaires designed to utilise three-alternative pick-one formats and case 3 BW formats.

4. Methods

4.1. Questionnaire design

The first part of the questionnaire provided information about the Pilbara region and current land uses, for example: cattle grazing; mining; and conservation areas. Questions on knowledge and experience of the region were interspersed with this background information. The second part of the questionnaire explained (iron ore) mine dewatering and possible management actions, including surplus groundwater uses. First, the consequences of disposing of surplus groundwater from mine operations along creeklines were explained. Using a set of three illustrations, developed with an ecologist, respondents were shown the expected consequences of creekline disposal of surplus groundwater over 20 years. The illustrations depict an initial shift from drought-tolerant species, to water-loving species and weeds, and finally to struggling dry-tolerant species competing with weeds.

Alternatives to creekline disposal of surplus groundwater were then described. One of these alternatives is to use the extracted groundwater to support or 're-water' local waterholes. As dewatering progresses, culturally significant waterholes can dry up. Once dry, these waterholes can no longer support water-dependent plants, animals, and birds. Many waterholes are important to local Aboriginal communities who have a cultural obligation to their ancestors and descendants to care for their Country and also to the wider Australian community.

Other potential uses of surplus groundwater were described. Respondents were told that over the past 50 years groundwater pumping from town bores had lowered the water table in regional centres, such as Tom Price, by around 30 metres and that groundwater levels would continue to fall in the future. To offset these water table declines and augment town water supplies, surplus groundwater can be injected into confined aquifers. An illustration depicted aquifer injection, with an explanation that between 20 and 60 years of additional town water supply was possible. A final irrigation option for using surplus groundwater was described, whereby intensive hay production would be used to offset the ecological effects of extensive cattle grazing. This option would lead to reduced cattle stocking rates on grazing land, which when combined with management actions (e.g. fencing, feral animal control) would help restore grazing land to habitat for native flora and fauna.

Based on these descriptions and the associated DCE attributes, respondents were asked to make choices among the status quo and two management alternatives. These alternatives were presented as different combinations of restored grazing land, preserved waterholes, and years of additional town water supply. A complete list of the attributes for surplus groundwater utilisation is shown in Table 1.

The third part of the questionnaire consisted of the choice tasks, debriefing questions and questions on environmental attitudes and socio-demographics. An example of a choice task is shown in Figure 2.

4.1.1. Elicitation question and payment scenario

The questionnaire indicates that creekline disposal of surplus groundwater is the mining industry standard in the Pilbara and is compliant with all mining permits and licences. It was further explained that other options for surplus groundwater disposal are not the responsibility of mining companies. This established the rationale for the involvement of the government, the use of public funds and the logic for the WTP framing of the choice questions. The payment vehicle was specified as a household levy that would be incurred each year for five years. Respondents were reminded of their budget constraints, and a short 'cheap talk script' following Morrison and Brown (2009) was used to

encourage respondents to make choices based on their true preferences. To minimise the potential for order effects, the non-cost attributes were randomised across respondents. Household cost always appeared in the last row of the choice task.

Table 1 Attributes, description and attribute levels

Attribute	Status Quo Level	Options B and C Levels
WATERHOLE, Preserve culturally-important waterholes	No natural waterholes	Waterholes: No natural waterholes remain; Preserve 1, 2, 3, 4, 5 natural waterholes
BIODIVERSE, Restore biodiverse grazing land	120,000 hectares degraded	Hectares restored: No Ha, 15,000, 30,000, 45,000, 60,000, 75,000
TOWN SUPPLY, Increase water supply for towns	Groundwater supply falling	Years of additional town water supply: 0, 20, 40, 60
COST, Levy per year for 5 years to your household	No Additional cost	Cost: \$25, \$50, \$75, \$100, \$125, \$150

Please read each of these questions carefully. Each option is a package to compare to the other options. Your answers will help determine the best use of this surplus groundwater. If these were the only three options available to you, which option would you vote for?




Features		Option A Maintain Current Situation	Option B Use water to:	Option C Use water to:
Culturally important waterholes		No natural waterholes remain	Preserve 5 natural waterholes	Preserve 3 natural waterholes
Water supply for towns		Groundwater supply falling	Supply 60 additional years of water	Supply 20 additional years of water
Grazing land		120,000 hectares degraded	Restore 15,000 hectares	Restore 75,000 hectares
Household cost <i>Per year for 5 years</i>	\$	\$0	\$50	\$100
I would vote in a referendum for: <i>Click on one box only</i>				
		Option A <input type="checkbox"/>	Option B <input type="checkbox"/>	Option C <input type="checkbox"/>
I like this option MOST: <i>Click on one box only</i>				
		Option A <input type="checkbox"/>	Option B <input type="checkbox"/>	Option C <input type="checkbox"/>
I like this option the LEAST: <i>Click on one box only</i>				
		Option A <input type="checkbox"/>	Option B <input type="checkbox"/>	Option C <input type="checkbox"/>

Figure 2 Example of a choice task

4.1.2. Focus groups, pre-testing, and elicitation format

Five focus groups (two in Adelaide, South Australia; one in Perth, Western Australia; and two in Sydney, New South Wales) were conducted to test survey language, alternative response formats, specifications of choice alternatives, attributes, and attribute levels. A few of the focus group participants indicated that they preferred answering BW over pick-one questions, and that they viewed a choice task posed as a pick-one referendum vote (referendum, hereafter) to be different from asking which alternative they liked best (or most). This focus group input provided the impetus for the split-sample design and associated hypotheses tested here.

Survey respondents were randomly assigned to one of six treatment conditions. Our analysis relies on data from three of the resulting six treatment conditions, denoted Conditions 1 - 3, and the other treatment conditions were designed for purposes unrelated to the proposed analysis. The choice task consisted of three alternatives, denoted Maintain Current Situation, Option B, and Option C

(Figure 2). After viewing these alternatives, Condition 1 respondents were asked which one they would vote for in a referendum (pick-one format). The pick-one referendum row disappeared from the screen once a choice was made. Then, two new rows appeared for Condition 1 respondents to elicit the best alternative and the worst alternative, thereby generating a complete preference ordering. Condition 2 consisted of only a pick-one referendum. Condition 3 consisted only of a BW format. The number of alternatives and attributes is the same across Conditions 1 - 3.

4.2. Survey sample

The survey was administered by the Australian panel provider, Online Research Unit (ORU). Potential respondents, stratified by age, gender, metro/rural and state of residence, were randomly selected from the ORU panel. The ORU panel consists of 300,000 community members who have agreed to participate in surveys for credits for shopping gift cards and entry into regular cash prize draws. The panel is regularly refreshed through online and offline techniques (for more detail see <http://theoru.com/>). Potential respondents ($n = 28,000$) were sent an initial email invitation to participate in a national survey. The invitations were sent out in waves ($n = 7,000$) for a series of treatment conditions. No information on the survey topic was provided in the invitation to avoid self-selection based on concern for environmental causes, groundwater, or mining interests. Up to two reminders to participate were sent. Once respondents completed informed consent, they were randomly assigned to different conditions. Details of the resulting sample are described in Section 5.

4.3. Experimental design

The final design is Bayesian D-efficient generated under the assumption of normally distributed prior parameters using a research version of the NGENE software.¹ Priors for the final design are the mean values and the standard errors as the standard deviation parameters from a pilot (see Bliemer and Collins (2016) for a discussion of the procedure followed in generating the design). The designs were generated using a co-ordinate exchange algorithm (Meyer and Nachtsheim 1995), minimising the Derror. This approach is commonly used in the literature (Scarpa and Rose, 2008) as it allows the researcher to avoid dominant alternatives in the choice task, as well as reducing sample size requirements (Rose and Bliemer, 2009).

The final design allows for all main effects and was constructed to allow for BW choices via the construction of pseudo-worst observations. In generating the design, it was assumed that the alternative chosen as best was deleted when constructing the pseudo-worst choice task. See Rose and Bliemer (2013) for a more detailed discussion of the design construction process for BW designs. The same design was used for the pick-one choice tasks. The design had 60 choice tasks and was blocked into 10 blocks of six choice tasks. A blocking column was generated by minimising the maximum absolute value of the correlation between the blocking column and the design attributes. The D-error of the final design was 0.000113.

4.4. Comparisons across elicitation formats

Initial hypothesis tests focus on the difference between parameter (here, WTP) estimates and error variances across the three conditions, or treatments. This follows approaches used in prior comparisons by Petrolia et al. (2018), one of the only comparisons of this type in the environmental DCE literature. We then further explore the potential for preference asymmetries related to framing differences across best and worst questions. As described by Johnston and Swallow (1999) for the

¹ The research version has the capability of generating the best-worst designs, but this feature has not been made available within the publicly available release version at the time of writing.

case of CV data, preference asymmetries of this type are inconsistent with the assumption of a single, fixed preference function, and hence with neoclassical welfare estimation. Several authors have noted that best and worst responses may represent different response frames, leading to different behavioural data generation processes that potentially cause preference and scale differences between the two response mechanisms (see, e.g. Rose 2013; Giergiczny et al. 2014). If such differences are present, then naïve pooling of the BW data will result in biased estimates due to data aggregation issues. The potential prevalence and impact of asymmetries of this type within environmental valuation DCEs are largely unknown.

The final set of evaluations diagnose symptoms of serial non-trading behaviours that emerged during the preference asymmetry tests described above, but that may be obscured when BW data are pooled. Combined with the framing effects discussed above, these behaviours can jeopardise the validity of inferences drawn from pooled BW models.

4.5. Econometric specification

The econometric specification is designed to accommodate possible scale and preference differences between the BW responses, as well as differences between treatment conditions. Whilst a number of methods exist that allow for the untangling of scale and preference differences in modelling (e.g. Bradley and Daly 1991), we have chosen to estimate models for each condition directly in WTP space (e.g. Train and Weeks 2005; Sonier et al. 2007), with separate models estimated for the separate response types. In doing so, we are first able to directly compare the model outputs across the different models, and secondly allow for correlated WTP distributions and random scale (see Scarpa et al. 2008).

For models estimated in WTP space, the coefficients represent marginal rate of substitution (MRS) distributions, and hence, it is not necessary to take the ratio of two coefficients in order to calculate marginal welfare effects. To illustrate the model, we first specify the utility function as separable in price, p , with the remaining k non-payment attributes designated as x_{nsjk} . Let U_{nsj} denote the utility of alternative j obtained by respondent n in choice situation s , which can be written as:

$$U_{nsj} = -\beta_{np} p_{nsj} + \sum_{k=1}^K \beta_{nk} x_{nsjk} + \varepsilon_{nsj}, \quad (1)$$

Given ordinal utility, it is possible to divide Equation (1) by the scale parameter, λ_n , which is inversely related to the error variance ε_{nsj} . This does not affect how behaviour is described by the model, but ensures that error variances are invariant over respondents, leading to:

$$U_{nsj} = \left(-\beta_{np}/\lambda_n\right) p_{nsj} + \sum_{k=1}^K \left(\beta_{nk}/\lambda_n\right) x_{nsjk} + \varepsilon_{nsj}, \quad (2)$$

Where ε_{nsj} is distributed multivariate normal with constant variance. The utility function may now be defined using $\theta_n = \left(-\beta_{np}/\lambda_n\right)$ and $\delta_n = \left(\beta_{nk}/\lambda_n\right)$ such that:

$$U_{nsj} = -\theta_n p_{nsj} + \sum_{k=1}^K \delta_n x_{nsjk} + \varepsilon_{nsj}. \quad (3)$$

The specification described by Equation (3) parameterises preferences in preference space. By re-specifying the utility function as follows, however, the model is estimated in WTP space, such that the resulting parameters ω_n are WTP rather than preference parameters:

$$U_{nsj} = -\theta_n p_{nsj} + \sum_{k=1}^K \theta_n \omega_n x_{nsjk} + \varepsilon_{nsj}. \quad (4)$$

That is, ω_n reflect the MRS between non-monetary attributes and money cost (or net income). This is the standard WTP space specification of the type illustrated by Train and Weeks (2005) and Scarpa et al. (2008).

The utility specification described by Equation (4) can be estimated using any form of discrete choice model, including a mixed multinomial logit model specification. The MRS parameters, ω_n in Equation (4), can therefore be specified as randomly distributed over the population. Given that the parameters ω_n are estimated directly, and hence, no ratios are required, any distribution of ω_n can be assumed. Here, we assume randomly distributed parameters for all the attributes, further allowing for the addition an error component associated with the two non-status quo alternatives (see Scarpa et al. 2005). The resulting utility function with error component is as follows:

$$U_{nsj} = -\theta_n p_{nsj} + \theta_n \left(\sum_{k=1}^K \omega_n x_{nsjk} + \eta_n d_j \right) + \varepsilon_{nsj}, \quad (5)$$

where η_n is normally distributed with zero mean and d_j is a dummy variable equal to one for the non-status quo alternatives, and zero for the status quo alternative.

All models assume that the price coefficient is log-normally distributed, with the remaining parameters being normally distributed. Further, all models allow for estimation of the full Cholesky matrix between the random parameter estimates (including the error component) (see Scarpa et al. 2008). Models are estimated using Python Biogeme (Bierlaire 2016) using 2000 Modified Latin Hypercube Sampling (MLHS) draws to obtain the simulated maximum-likelihood estimates (see Hensher et al. 2015 for a description of MLHS draws).

4. Results and discussion

The survey was implemented online during September - October 2013, with a response rate of 12.6% across six treatment conditions. All other results in this section refer only to the three conditions (BW and referendum, BW only, and referendum only). The socio-demographic characteristics of the respondents in the three conditions are summarised in Table 2. The sample of 1,408 is older, more educated, and has a higher income than the general Australian population (ABS 2013). For example, in 2013 the median household income of the Australian population was \$64,200, whereas 56.8% of the survey respondents had a household income level over \$65,000. The sample has a higher proportion of households with a degree or higher (34.8%), compared to the population (23.8%). Modest differences in this type between the characteristics of stated preference survey samples and general populations are common in the literature.

Table 2 Sample characteristics

Variable	
Mean age, years	46.8
Median age, years	47
Mean household size, person	2.9
Proportion female	48.8%
Aboriginal or Torrens Strait Islander	1.8%
Income categories	
under \$31,149	26.9%
\$31, 150 to \$64, 949	27.6%
\$64, 950 to \$103, 949	24.1%
\$103, 950 to \$155, 949	14.4%
\$155, 950 +	7.0%
Highest education levels obtained	
less than year 12	12.0%
Year 12	16.4%
TAFE cert/diploma	34.5%
Bachelor's degree	20.3%
Grad diploma/Post-Graduate Degree	14.5%

A total 594 respondents were assigned to and completed Condition 1 tasks (referendum followed by BW). Because respondents were given six tasks to complete, this represents 3,564 referendum choices, and 7,128 pooled BW observations. Four hundred and eight respondents were randomly assigned to Condition 2 completing six referendum questions each, and 406 respondents completed BW-only responses in Condition 3.2 Tables 3 and 4 present model results for all three conditions. For Condition 1, four models are presented in Table 3. Model 1a is estimated using data from the referendum response, whilst Model 1b and Model 1c are estimated on the best and the worst responses separately. Model 1d is estimated based on pooled BW data. In Condition 1, the option that the participant voted for in a referendum is closely associated with the option the participant liked the most. This does not mean that they are necessarily equivalent as the referendum was asked first and may have an impact on the best response.

Table 4 presents four models, where model 2a is based on data related to the referendum response collected as part of condition 2 and models 3a, b and c are estimated on the best, worst and pooled BW responses associated with condition 3. At the bottom of each table are the WTP estimates for the three non-cost attributes, based on the mean coefficients of the WTP distributions. Confidence intervals calculated via the Delta method are provided for each of these WTP estimates. Whilst it is possible to estimate confidence intervals for the entire WTP distribution, such intervals typically provide less than useful information, with very large confidence intervals (see Bliemer and Rose 2008).

Table 3 Condition 1 models

		1a: Referendum only		1b: Best only		1c: Worst only		1d: BW naïvely pooled	
Variable	Moment	Par.	(Rob. t-rat.)	Par.	(Rob. t-rat.)	Par.	(Rob. t-rat.)	Par.	(Rob. t-rat.)
Preserve Additional Waterhole	Mean	48.035	(10.43)	58.200	(4.80)	55.097	(131.76)	38.357	(9.65)
	Std Dev.*	3.136	(1.65)	9.573	(4.27)	5.032	(106.02)	3.419	(9.15)
Restore Grazing Land	Mean	1.990	(10.22)	2.920	(4.26)	3.070	(67.87)	1.879	(9.72)
	Std Dev.*	2.002	(2.44)	4.168	(4.25)	4.127	(10.54)	1.590	(6.83)
Additional Year of Water Supply	Mean	2.975	(10.48)	4.610	(4.87)	2.896	(15.93)	2.183	(9.13)
	Std Dev.*	3.170	(2.07)	9.446	(4.07)	3.100	(9.86)	2.176	(8.27)
Cost	Mean	0.153	(1.09)	-0.252	(-1.21)	2.769	(66.97)	0.098	(0.87)
	Std Dev.*	2.056	(12.77)	0.115	(0.71)	9.230	(452.45)	1.520	(7.68)
<i>Error Component</i>									
Error Component*		5.154	(2.80)	6.098	(4.88)	3.298	(219.95)	4.258	(9.40)
<i>Model fit</i>									
Respondents		594							
Observations		3564		3564		3564		7128	
LL(0)		-3915.454		-3915.454		-2470.377		-6385.831	
LL(β)		-2451.63		-2120.031		-1441.996		-3839.574	
ρ^2		0.374		0.459		0.416		0.399	
adj. ρ^2		0.371		0.456		0.413		0.396	
<i>Willingness to Pay Confidence intervals at mean of random parameter distribution</i>									
Preserve Additional Waterhole		\$48.04 [\$39.01 - \$57.05]		\$58.20 [\$34.48 - \$81.91]		\$55.10 [\$54.27 - \$55.91]		\$38.36 [\$30.55 - \$46.15]	
Restore 1000 ha Grazing Land		\$1.99 [\$1.60 - \$2.37]		\$2.92 [\$1.57 - \$4.26]		\$3.07 [\$2.98 - \$3.15]		\$1.88 [\$1.50 - \$2.25]	
Additional Year of Water Supply		\$2.98 [\$2.41 - \$3.53]		\$4.61 [\$2.75 - \$6.46]		\$2.90 [\$2.839 - \$4.66]		\$2.18 [\$1.71 - \$2.65]	

*Standard errors computed using Delta method based on the full Cholesky matrix; the full Cholesky matrix is available upon request for each model.

Table 4 Condition 2 and 3 models

		Condition 2		Condition 3					
		2a: Referendum only		3a: Best only		3b: Worst only		3c: BW naïvely pooled	
Variable	Moment	Par.	(Rob. t-rat.)	Par.	(Rob. t-rat.)	Par.	(Rob. t-rat.)	Par.	(Rob. t-rat.)
Preserve Additional Waterhole	Mean	33.032	(8.87)	24.714	(9.39)	100.806	(14.48)	24.449	(8.84)
	Std Dev.*	2.480	(6.57)	2.900	(10.96)	22.172	(2.61)	2.072	(11.81)
Restore Grazing Land	Mean	1.528	(9.04)	1.354	(10.25)	-10.162	(-64.82)	0.102	(1.10)
	Std Dev.*	0.973	(6.11)	1.178	(9.12)	10.097	(89.52)	0.972	(7.48)
Additional Yr of Water Supply	Mean	1.088	(9.31)	2.121	(11.40)	3.150	(5.10)	2.000	(8.86)
	Std Dev.*	1.236	(8.33)	1.830	(9.71)	1.568	(69.67)	1.242	(6.72)
Cost	Mean	1.104	(4.61)	0.954	(4.82)	-2.720	(-3.20)	0.108	(1.08)
	Std Dev.*	1.303	(5.98)	1.439	(6.54)	3.500	(2.61)	1.509	(11.99)
<i>Error Component</i>									
Error Component*		2.602	(5.31)	4.605	(7.25)	91.32	(99.95)	8.08	(12.43)
<i>Model fit</i>									
Respondents		408		406					
Observations		2448		2436	2436		4872		
LL(0)		-2689.403		-2676.220		-1688.507		-4364.726	
LL(β)		-1591.587		-1598.176		-527.942		-2587.209	
ρ^2		0.408		0.403		0.687		0.407	
adj. ρ^2		0.404		0.398		0.685		0.403	
<i>Willingness to Pay Confidence intervals</i>									
Preserve Additional Waterhole		\$33.03 [\$25.74 - \$40.32]		\$24.71 [\$19.50 - \$29.92]		\$100.81 [\$-41.8 - \$243.4]		\$24.45 [\$19.01 - \$29.87]	
Restore 1000 ha Grazing Land		\$1.53 [\$1.23 - \$1.82]		\$1.35 [\$1.09 - \$1.61]		-\$10.16 [\$-13. - \$-6.6]		\$0.10 [\$-0.0 - \$0.28]	
Additional Year of Water Supply		\$1.09 [\$0.76 - \$1.40]		\$2.12 [\$1.75 - \$2.48]		\$3.15 [\$-8.96 - \$14.2]		\$2.00 [\$1.53 - \$2.46]	

*Standard errors computed using Delta method based on the full Cholesky matrix; the full Cholesky matrix is available upon request for each model.

None of the models include an alternative specific constant (ASC) for the status quo alternative. The inclusion of ASCs tended to cause problems with the estimation of other parameters within the model. Before discussing all model results in detail, we note that the models based only on the worst response data (models 1c and 3b) have either very large parameter estimates, or t-ratios (e.g. the t-ratio for the standard deviation of the cost parameter is 452.45 for Model 1c whilst the error component for model 3b is 99.95). This suggests the presence of significant problems with these data including estimating the Hessian matrix of these models. These results suggest that the two worst-response data sets are ill-conditioned (we discuss why this might be the case below). Further, testing for differences in scale and preferences suggests preference asymmetries across best and worst responses (Rose 2013) in both Conditions 1 and 3. This suggests the best and worst data should not be pooled. Thus, presentation of results for such 'naïvely' pooled BW data is done in order to highlight the contrast between the seemingly reasonable results of the pooled data and the results from modelling the worst data.

For all the models in Tables 3 and 4, the error components associated with the non-status quo options are positive and significant. For all the models in Table 3 except 1c in Condition 1, the means of the cost random parameters are not statistically significant. However, this does not imply an average marginal utility of zero for cost (or a zero scale), recalling that the mean of a log-normal distribution is $e^{\mu + \frac{1}{2}\sigma^2}$.

Model 1a in Table 3 and Condition 2 in Table 4 are estimated from referendum data. WTP coefficients for each of the groundwater management options in model 1a and Condition 2 are all statistically significant. In the case of waterholes and grazing land, the confidence intervals on Model 1a and condition 2 overlap. However, a Poe test (Poe et al. 2005) indicates that the estimates are not statistically different. That is, we cannot reject the null hypothesis of equal WTP estimates across these two referendum treatments. The WTP estimates from Condition 2 imply that our sample of Australians is willing to pay \$33.03 [95% CI: \$25.74 to \$40.32] per year to preserve an additional waterhole; \$1.53 [95% CI: \$1.23 to \$1.82] per year to restore an additional 1,000 ha of grazing land to habitat area; and \$1.09 [95% CI: \$0.76 to \$1.40] per year to extend town water supply for 1 year.

In Table 3, a comparison of mean WTP estimates from the Condition 1 referendum and naïvely pooled BW data might lead to an erroneous conclusion that the two formats yield the same WTP estimates for all three attributes. A similar comparison of Condition 2 and naïvely pooled BW data in Condition 3 (Table 4) might lead to a similar erroneous conclusion for preserving waterholes (but not for the other two attributes). However, the differences between best and worst models in both cases (1b versus 1c; 3a versus 3b), together with the ill-conditioned worst data (see additional discussion below), should dissuade any such comparisons. Results such as these suggest that caution should be exercised when drawing conclusions from pooled BW data, without first examining whether pooling is justified. These problems can persist even when naïvely pooled results appear satisfactory from a superficial perspective (as shown here in Tables 3 and 4).

5.1. Diagnosing problems in best-worst data

To further explore patterns in the worst data, Table 5 presents the mean, median, and standard deviation for the log-normally distributed cost/scale parameter for each model and data subset. Results suggest severe scale issues with the worst data, particularly for (but not limited to) the Condition 1 data set. These are illustrated by extreme values for the mean and standard deviation within the worst data, compared to other formats.

Table 5 Population moments of scale/cost parameter

	1a: Referendum only	1b: Best only	1c: Worst only	1d: BW Naïvely combined	2a: Referendum only	3a: Best only	3b: Worst only	3c: BW naïvely pooled
Mean	9.65	0.78	5.03×10^{19}	3.50	7.05	7.31	30.11	3.48
Median	1.17	0.78	15.94	1.10	3.02	2.60	0.07	1.11
Std Dev.	79.26	0.09	1.59×10^{38}	10.55	14.89	19.25	13766.56	10.29

Table 6 Number of non-trading respondents by choice by treatment condition

Choice	Non-trading	Reject	Retain	Total
<i>Condition 1</i>				
Referendum	Non-Trader all alts	94 (15.82%)	500 (84.18%)	594 (100.00%)
Referendum	Non-trader SQ	67 (11.28%)	527 (88.72%)	594 (100.00%)
Best	Non-Trader all alts	87 (14.65%)	507 (85.35%)	594 (100.00%)
Best	Non-trader SQ	58 (9.76%)	536 (90.24%)	594 (100.00%)
Worst	Non-Trader all alts	313 (52.69%)	281 (47.31%)	594 (100.00%)
Worst	Non-trader SQ	291 (48.99%)	303 (51.01%)	594 (100.00%)
All choices	Non-Trader all alts	380 (63.97%)	214 (36.03%)	594 (100.00%)
All choices	Non-trader SQ	364 (61.28%)	230 (38.72%)	594 (100.00%)
<i>Condition 2</i>				
Referendum	Non-Trader all alts	21 (5.15%)	387 (94.85%)	408 (100.00%)
Referendum	Non-trader SQ	14 (3.43%)	394 (96.57%)	408 (100.00%)
<i>Condition 3</i>				
Best	Non-Trader all alts	65 (16.01%)	341 (83.99%)	406 (100.00%)
Best	Non-trader SQ	54 (13.30%)	352 (86.70%)	406 (100.00%)
Worst	Non-Trader all alts	204 (50.25%)	202 (49.75%)	406 (100.00%)
Worst	Non-trader SQ	192 (47.29%)	214 (52.71%)	406 (100.00%)
All choices	Non-Trader all alts	251 (61.82%)	155 (38.18%)	406 (100.00%)
All choices	Non-trader SQ	246 (60.59%)	160 (39.41%)	406 (100.00%)

To better understand why such patterns occur (and why these data are problematic), Table 6 illustrates the extent of ‘non-traders’ found within each treatment condition by choice type (i.e. response format). We distinguish between two types of non-trading behaviour. The first represents respondents who always select the same alternative over all six choice tasks (non-trader all alts). The second are respondents who always select the status quo alternative in all six tasks (non-trader SQ). As such, the latter represent a subset of the more general non-trading respondents (i.e. non-trader all alts). Within the table, the ‘reject’ column identifies those for which trading behaviour is rejected (i.e. non-traders). In contrast, the ‘retain’ column identifies those who do not always choose the same alternative, that is illustrate trading behaviour. For example, for the referendum choice question in Condition 1, 94 (15.82 per cent) of respondents were observed to select the same alternative over all six tasks. Of these 94 respondents, 67 (11.28 per cent of the entire sample) always selected the status quo alternative.

For Conditions 1 and 3, Table 6 also presents data on the number of respondents who display non-trading data for at least one response format. For example, out of the 594 respondents assigned to Condition 1, 380 (63.97 per cent) respondents selected the same alternative over all six tasks for either the referendum, best, or the worst choice question, or some combination of all three response formats. Of these respondents, 364 (61.28 per cent) always selected the status quo alternative in at least one of the response formats.

Such sequences of non-trading choice behaviour may occur due to valid preference structures. However, identical behaviour may also emerge due to heuristics or other response patterns that are inconsistent with neoclassical, fully compensatory choice processes. Hence, whether to include or exclude such responses in modelling efforts is not obvious. It is clear that such patterns are most common in the worst data in Conditions 1 and 3 and there is a significant representation of non-trading related to the status quo alternative. A large proportion of respondents are universally choosing the status quo as worst.

A pattern such as this could reflect an anti-status quo or other form of strategic bias that emerges only within the negative preference domain of worst responses. That is, the negative framing of worst-alternative elicitation might invite symbolic, non-compensatory responses by a subset of respondents who wish to protest against the status quo, regardless of the choice attributes. These respondents hence always choose the status quo as worst.

Alternatively, the same behaviour might emerge due to insufficiently high bid levels in the experimental design, similar to the well-known 'fat tails' problem in binary choice CV (Ready and Hu 1995). It is common in valuation DCEs for environmental attribute levels of the non-status quo alternatives to represent outcomes viewed as improvements over the status quo (i.e. to have positive marginal utility). As a result, the change in utility related to the combined non-cost attributes is positive within all non-status quo alternatives, compared to the status quo. If the cost of the non-status quo alternatives (the bid level) is insufficient to offset this utility gain, then the status quo will always be seen as the least desired outcome. Hence, a design with insufficiently high bid levels in BW elicitation can lead to serial choices wherein the status quo is always selected as the worst alternative. In this case, non-trading choices are a legitimate reflection of preferences under a bid design with levels that are insufficient to invoke trading behaviour.

Here, we cannot disentangle serial non-status quo choices due to non-compensatory or strategic response patterns from those due to bid design effects. Another possibility is that respondents might have always chosen the first (left-hand) alternative as worst. Although this is possible in concept, no evidence of this behaviour emerged from focus groups, and there is little evidence in the environmental valuation DCE literature for such extreme left-versus-right choice patterns. Hence, we view this to be an unlikely explanation for the observed response data. Though the levy within the bid design extended up to \$150 per year for 5 years (a seemingly high level for this type of environmental change), it is possible that the range of the cost attribute may not have been large enough to prevent non-trading. If this is the case, respondents would be expected to select one of the two non-status quo alternatives as their most preferred alternative in the first instance, with the status quo being selected as the worst choice in the second instance. The status quo alternative will appear to be non-traded over all choice tasks for the worst choice question, even if trading between the two non-status quo alternatives was observed for the best choice questions.

Regardless of the cause, the presence of significant number of non-traders within a dataset may cause modelling issues, particularly for ASCs (e.g. Rose 2013), as observed in the results presented here. If these response patterns are caused by the bid design, they could also manifest in the referendum data via a pattern in which many respondents never choose the status quo, over all choice tasks (i.e. they always choose either Option A or Option B). This is not serial choice behaviour of the type typically discussed in the environmental DCE literature, but can also lead to difficulties with estimation, particularly related to the status quo ASC. Non-trading patterns of this type may also cause inconsistencies across best and worst data that, when combined with any preference asymmetries that may exist, preclude valid pooling of BW data. These non-trivial data inconsistencies can belie the seemingly reasonable results of pooled BW models and suggest that careful exploration of the data should precede the use and interpretation of pooled BW model results is required.

5.2. Comparison of condition 2 results to the literature

The WTP estimates of Condition 2 were estimated from a DCE which did not involve BW elicitation, and hence can be discussed separately. This section presents these results briefly and contrasts them to findings of prior work, to supplement the methodological findings presented above. The

purpose of doing so is to support potential policy analysis and future benefit transfers (Brouwer et al. 2015), recognising that there are only a few Australian studies that explore WTP with methods similar to those used in our case study: Aboriginal waterholes; rangeland restoration; and town water supply. Unless otherwise noted, all WTP estimates are interpreted as annual values over a five-year payment horizon.

Condition 2 results suggest that respondents are willing to pay \$33.03 per household to preserve an additional cultural important waterhole. Using a discount rate of 3%, this amounts to a lump-sum equivalent of \$152 and is comparable to the findings of Zander et al. (2010) who consider WTP to improve the condition of billabongs (waterholes) of importance to Aboriginal people in northern Australia. Smaller values were found by Gillespie and Kragt (2012) in the context of mitigating stream impacts resulting from coal mining. They use an online referendum DCE and estimate the WTP to protect Aboriginal sites at \$0.27 per person per site. In contrast, Rolfe and Windle (2003) report that a non-Aboriginal sample did not value high levels of site protection (e.g. negative WTP) for Aboriginal cultural heritage sites in the Fitzroy Basin in Queensland, Australia. Respondents were more concerned with environmental issues and economic development.

For rangeland restoration, respondents had a lump-sum equivalent WTP of \$7.05 per 1,000 ha restored. Although some broadly comparable estimates are available from other countries (e.g. Dissanayake and Ando 2014 in the US), no directly comparable WTP study could be located that elicited value of destocking land and restoring arid rangeland habitat to support a diversity of flora and fauna in Australia. Similar resource types are studied by Greiner (2016), who reports a mean willingness to accept compensation among pastoralists of \$11.08/ha for a conservation strategy of complete exclusion of cattle and \$3.45/ha for land being spelled every year for an extended time.

For an additional year of town water supply, respondents had a lump-sum equivalent WTP of \$5.02 per household. Although we could identify no other studies that estimate the WTP to extend town water supply, there are studies that report WTP to reduce domestic water restrictions. For example, Brouwer et al. (2015) find that WTP for reducing domestic water restrictions in an Australian case study (Fitzroy basin, Queensland) was not significantly different from zero. In contrast, in the coal mining town of Moranbah, central Queensland, Ivanov and Rolfe (2011) estimated households WTP to avoid any restriction on indoor domestic use at \$218. These combined results suggest that values related to household water supplies vary considerably across Australian case studies.

6. Conclusions

There has been increasing interest in the use of BW elicitation across different areas of application such as health, environmental valuation, marketing, and transportation because BW may be an efficient means of gathering preference information. Our initial interest in designing this set of discrete choice experiments was the potential capacity of BW methods to elicit robust preference and welfare estimates compared to other elicitation formats such as pick-one DCE. Our results suggest that caution is warranted beginning with whether the best and worst data can be pooled, let alone compared to pick-one DCE. A second problem is that we find strong evidence of non-trading, or anti status quo behaviour, in the worst-response data. We offer two alternative possible reasons. First, there may be an inherent tendency for non-trading to occur in worst-response framing in the presence of a status quo option. Second, non-trading may be the result of an insufficient high bid design, though we are sceptical of this explanation. Perhaps even a combination of these two factors may be the cause. Regardless, these results suggest that practitioners need to check for preference asymmetry and non-trading behaviour. 'Naïvely' pooling best and worst data may disguise a host of problems.

These are one set of results conducted under a particular set of conditions. We believe there is a need for additional well-executed, independent tests to determine whether preference asymmetry and non-trading is a wide-spread problem.

Acknowledgments

We would like to acknowledge the funding provided by CSIRO Water for a Healthy Country Research Flagship, Rio Tinto Iron Ore, and International Union of Conservation of Nature. However, the research decisions taken in this article are those of the authors. The assistance of Martin Nolan in the creation of spatial calculations and map, and Darran King in data handling is also gratefully acknowledged.

References

- ABC News (2013). Mining fueled Agriculture. <http://www.abc.net.au/news/rural/2013-04-22/mining-fuelled-agriculture/4640406#&gclid=1&pxml:id=1> [accessed 20 June 2018].
- Australian Bureau of Statistics (2013). 2011 census quickstats. http://www.censusdata.abs.gov.au/census_services/getproduct/census/2011/quickstat/0 [accessed 20 June 2018].
- Australian Bureau of Statistics (2015). Mining. Year Book Australia 2012. <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1301.0Main+Features292012> [accessed 7 July 2018].
- Australian Bureau of Statistics (2018). 8155.0 - Australian Industry, 2016-17. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/8155.0> [accessed 20 June 2018].
- Beck, M., Rose, J.M. and Greaves, S. (2017). I can't believe your attitude: a joint estimation of best worst attitudes and electric vehicle choice, *Transportation* 44, 753-772.
- Bliemer, M.C.J. and Rose, J.M. (2008). Confidence intervals of willingness-to-pay for random coefficient logit models, *Transportation Research Part B* 58, 199-214.
- Bierlaire, M. (2016). Pythonbiogeme: a short introduction. Report TRANSP-OR 160706, Series on Biogeme. Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Federale de Lausanne, Switzerland.
- Bliemer, M.C.J. and Collins, A.T. (2016). On determining priors for the generation of efficient stated choice experimental designs, *Journal of Choice Modelling* 21, 10-14.
- Boyle, K.J. (2017). Contingent valuation in practice, in: Champ, P.A., Boyle, K.J. and Brown, T.C. (eds), *A Primer in Nonmarket Valuation*, 2nd edn. Springer, Dordrecht, The Netherlands, pp. 83-131.
- Bradley, M. and Daly, A. (1991). Estimation of logit choice models using mixed stated and revealed preference information, *Proceedings of 6th International Conference on Travel Behaviour*, 117-133
- Brefle, W.S. and Rowe, R.D. (2002). Comparing choice question formats for evaluating natural resource tradeoffs, *Land Economics* 78, 298-314.
- Brouwer, R., Martin-Ortega, J., Dekker, T., Sardonini, L., Andreu, J., Kontogianni, A., Skourtos, M., Raggi, M., Viaggi, D., Pulido-Velazquez, M., Rolfe, J. and Windle, J. (2015). Improving value transfer through socio-economic adjustments in a multicountry choice experiment of water conservation alternatives, *Australian Journal of Agricultural and Resources Economics* 59, 458-478.
- Carson, R.T. and Groves, T. (2007). Incentive and informational properties of preference questions, *Environmental and Resource Economics* 37, 181-210.

- Collins, A.C. and Rose, J.M. (2011). Estimation of stochastic scale with best-worst data, International Choice Modelling Conference, July 4–6. Leeds, UK.
- Collins, J.P. and Vossler, C.A. (2009). Incentive compatibility tests of choice experiment value elicitation questions, *Journal of Environmental Economics and Management* 58, 226–235.
- Currell, M. (2015). Groundwater: the natural wonder that needs protecting from coal seam gas. The Conversation. <https://theconversation.com/groundwater-the-natural-wonder-that-needs-protecting-from-coal-seam-gas-41978> [accessed 7 July, 2018].
- DeShazo, J.R. and Fermo, G. (2002). Designing choice sets for stated preference methods: the effects of complexity on choice consistency, *Journal of Environmental Economics and Management* 44, 123–143.
- Dissanayake, S.T.M. and Ando, A.W. (2014). Valuing grassland restoration: proximity to substitutes and trade-offs among conservation attributes, *Land Economics* 90, 237–259.
- Dyachenko, T., Walker Reczek, R. and Allenby, G.M. (2014). Models of sequential evaluation in best-worst choice tasks, *Marketing Science* 33, 828–848.
- Flynn, T.N., Louviere, J.J., Peters, T.J. and Coast, J. (2007). best-worst scaling: What it can do for health care research and how to do it, *Journal of Health Economics* 26, 171–189.
- Giergiczny, M., Hess, S., Dekker, T. and Chintakayala, P.K. (2014). Testing the consistency (or lack thereof) between choices in best-worst surveys. Transportation Research Board Annual Meeting Paper #14-4858. TRB 93rd Annual Meeting Compendium of Papers.
- Gillespie, R. and Kragt, M. (2012). Accounting for nonmarket impacts in a benefit-cost analysis of underground coal mining in New South Wales, Australia, *Journal of Benefit-Cost Analysis* 3, 1–29
- Government of Western Australia (2000). Water Quality Protection Guidelines No. 11. Mining and Mineral Processing: Mine dewatering. Department of Minerals and Energy Western Australia and Department of Water and Environmental Regulation. http://www.water.wa.gov.au/__data/assets/pdf_file/0014/4370/44630.pdf [accessed 20 June 2018].
- Greiner, R. (2016). Factors influencing farmers' participation in contractual biodiversity conservation: a choice experiment with northern Australian pastoralists, *Australian Journal of Agricultural and Resource Economics* 60, 1–21.
- Hanley, N., Mourato, S. and Wright, R.E. (2001). Choice modelling approaches: a superior alternative for environmental valuation? *Journal of Economic Surveys* 15, 435–462.
- Hensher, D.A. (2006). Revealing differences in willingness to pay due to the dimensionality of stated choice designs: an initial assessment, *Environmental and Resource Economics* 34, 7–44.
- Hensher, D.A., Rose, J.M. and Greene, W. (2015). *Applied Choice Analysis*, 2nd edn. Cambridge University Press, Cambridge, UK.
- Ivanova, G. and Rolfe, J. (2011). Assessing development options in mining communities using stated preference techniques, *Resources Policy* 36, 255–264.
- Johnston, R.J. and Swallow, S.K. (1999). Asymmetries in ordered strength of preference models: Implications of focus shift for discrete choice preference estimation, *Land Economics* 75, 295–310.

- Johnston, R.J., Boyle, K.J., Adamowicz, W., et al. (2017). Contemporary guidance for stated preference studies, *Journal of the Association of Environmental and Resource Economists* 4, 319–405.
- Louviere, J.J., Flynn, T.N. and Marley, A.A.J. (2015). *Best-worst Scaling: Theory, Methods and Applications*. Cambridge University Press, Cambridge, UK.
- Lusk, J.L. and Briggeman, B. (2009). Food Values, *American Journal of Agricultural Economics* 91, 184–196.
- Marley, A.A.J. and Louviere, J.J. (2005). Some probabilistic models of best, worst, and bestworst choices, *Journal of Mathematical Psychology* 49, 464–480.
- Meyer, R.K. and Nachtsheim, C.J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs, *Technometrics* 37, 60–69.
- Meyerhoff, J., Oehlmann, M. and Weller, P. (2015). The influence of design dimensions on stated choices in an environmental context, *Environmental and Resource Economics* 61, 385–407.
- Morrison, M. and Brown, T. (2009). Testing the effectiveness of certainty scales, cheap talk, and dissonance-minimization in reducing hypothetical bias in contingent valuation studies, *Environmental and Resource Economics* 44, 307–326.
- Petrolia, D.R., Interis, M.G. and Hwang, J. (2018). Single-choice, repeated-choice, and bestworst scaling elicitation formats: Do results differ and by how much? *Environmental and Resource Economics* 69, 365–393.
- Poe, G.L., Girard, K.L. and Loomis, J.B. (2005). Computational methods for measuring the difference of empirical distributions, *American Journal of Agricultural Economics* 87, 353–365.
- Ready, R.C. and Hu, D. (1995). Statistical approaches to the fat tail problem for dichotomous choice contingent valuation, *Land Economics* 71, 491–499.
- Rolfe, J. and Windle, J. (2003). Valuing the protection of aboriginal cultural heritage sites, *Economic Record* 79, S85–S95.
- Rose, J.M. (2013). *Interpreting discrete choice models based on best-worst data: A matter of framing*. Institute of Transport and Logistics Studies, The Australian Key Centre in Transport and Logistics Management, The University of Sydney. Working Paper ITLS-WP13-22.
- Rose, J.M. and Bliemer, M.C.J. (2009). Constructing Efficient Stated Choice Experimental Designs, *Transport Reviews* 29, 587–617.
- Rose, J.M. and Bliemer, M.C.J. (2013). Incorporating analyst uncertainty in model specification of respondent processing strategies into efficient designs for logit models, 59th World Statistics Congress, Hong Kong, 30th August 2013.
- Scarpa, R. and Rose, J.M. (2008). Designs efficiency for nonmarket valuation with choice modelling: how to measure it, what to report and why, *Australian Journal of Agricultural and Resource Economics* 52, 253–282.
- Scarpa, R., Ferrini, S. and Willis, K. (2005). Performance of error component models for status-quo effects in choice experiments, in Scarpa, R. and Alberini, A. (eds), *Applications of Simulations*

Methods in Environmental Resource Economics. Kluwer Academics Publisher, Dordrecht, The Netherlands, pp. 247–274.

Scarpa, R., Thiene, M. and Train, K. (2008). Utility in willingness to pay space: a tool to address confounding random scale effects in destination choice to the Alps, *American Journal of Agricultural Economics* 90, 994–1010.

Scarpa, R., Notaro, S., Louviere, J. and Rafelli, R. (2011). Exploring scale effects of best/worst rank ordered choice data to estimate visitors benefits of tourism in alpine grazing commons, *American Journal of Agricultural Economics* 93, 813–828.

Smith, V.K. (2006). Fifty years of contingent valuation, in Alberini, A. and Kahn, J.R. (eds), *Handbook on Contingent Valuation*. Edward Elgar Publishing, Northampton, MA, pp. 7–65.

Sonnier, G., Ainslie, A. and Otter, T. (2007). Heterogeneity distributions of willingness-to-pay in choice models, *Quantitative Marketing Economics* 5, 313–31.

Soto, J.R., Escobedo, F.J., Khachatryan, H. and Adams, D.C. (2018). Consumer demand for urban forest ecosystem services and disservices: examining trade-offs using choice experiments and best-worst scaling, *Ecosystem Services* 29, 31–39.

Swait, J.D. and Adamowicz, W.L. (2001). Choice environment, market complexity and consumer behaviour: a theoretical and empirical approach for incorporating decision complexity into models of consumer choice, *Organizational Behavior and Human Decision Processes* 86, 141–167.

Train, K. and Weeks, M. (2005). Discrete choice models in preference space and willingness to-pay space, in Scarpa, R. and Alberini, A. (eds), *Applications of Simulations Methods in Environmental Resource Economics*. Kluwer Academics Publisher, Dordrecht, The Netherlands, pp. 1–16.

Vermeulen, G.P. and Vandebroek, M. (2011). Rank-order conjoint experiments: efficiency and design, *Journal of Statistical Planning and Inference* 141, 2519–2531.

Vossler, C.A., Doyon, M. and Rondeau, D. (2012). Truth in consequentiality: Theory and field evidence on discrete choice experiments, *American Economic Journal: Microeconomics* 4, 145–171.

WA DOW (2012). Pilbara groundwater allocation plan: Looking after all our water needs. Draft Report, Department of Water, Water Resource Allocation and Planning Report series, October 2012. Government of Western Australia.

Yangui, A., Akaichi, F., Costa-Font, M. and Gil, J.M. (2019). Comparing results of ranking conjoint analyses, best-worst scaling and discrete choice experiments in a nonhypothetical context, *Australian Journal of Agricultural Resource Economics* 63, 221–246.

Zander, K.K., Garnett, S. and Straton, A. (2010). Trade-offs between development, culture and conservation-willingness to pay for tropical river management among urban Australians, *Journal of Environmental Management* 91, 2519–2528.