Netspar

Two for the price of one

If moving beyond traditional single-best discrete choice experiments, should we use best-worst, best-best or ranking for preference elicitation?

Samare Huls, Emily Lancsar, Bas Donkers, Jemimah Ride

NETSPAR ACADEMIC SERIES

Health Economics WILEY

Two for the price of one: If moving beyond traditional single-best discrete choice experiments, should we use best-worst, best-best or ranking for preference elicitation?

Samare P. I. Huls[1,2] ⓘ | **Emily Lancsar**[3] | **Bas Donkers**[2,4] ⓘ | **Jemimah Ride**[5] ⓘ

[1]Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, the Netherlands

[2]Erasmus Choice Modelling Centre, Erasmus University Rotterdam, Rotterdam, the Netherlands

[3]Department of Health Services Research and Policy, Research School of Population Health, Australian National University, Canberra, Australian Capital Territory, Australia

[4]Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, the Netherlands

[5]Health Economics Unit, Melbourne School of Population and Global Health, University of Melbourne, Parkville, Victoria, Australia

Correspondence
Samare P. I. Huls, Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam 1738, 3000, the Netherlands.
Email: huls@eshpm.eur.nl

Abstract

This study undertook a head-to-head comparison of best-worst, best-best and ranking discrete choice experiments (DCEs) to help decide which method to use if moving beyond traditional single-best DCEs. Respondents were randomized to one of three preference elicitation methods. Rank-ordered (exploded) mixed logit models and respondent-reported data were used to compare methods and first and second choices. First choices differed from second choices and preferences differed between elicitation methods, even beyond scale and scale dynamics. First choices of best-worst had good choice consistency, scale dynamics and statistical efficiency, but this method's second choices performed worst. Ranking performed best on respondent-reported difficulty and preference; best-best's second choices on statistical efficiency. All three preference elicitation methods improve efficiency of data collection relative to using first choices only. However, differences in preferences between first and second choices challenge moving beyond single-best DCE. If nevertheless doing so, best-best and ranking are preferred over best-worst DCE.

KEYWORDS
best-best data, best-worst scaling, choice modeling, discrete choice experiment, preference elicitation, rank-ordered data

JEL CLASSIFICATION
C350, C590, C830, I100

1 | INTRODUCTION

Discrete choice experiments (DCEs) are widely used to answer a range of research questions relating to health care preferences (de Bekker-Grob et al., 2012; Soekhai et al., 2019). More recently academic interest in the way in which respondents process information and in which DCE questions are asked has been increasing (Cornelsen et al., 2020; Genie et al., 2021; Gonzalez Sepulveda et al., 2021; Mansfield et al., 2020). In DCEs it is standard practice to elicit preferences by repeatedly presenting respondents with a range of alternatives presented in choice sets. In each choice set, respondents evaluate alternatives that differ in their attribute levels. These traditional DCEs only ask participants to choose the best alternative per set. In addition to asking respondents their single-best alternative, there is an increasing interest in obtaining more preference information per set by also asking participants to choose among the remaining alternatives, which can lead to elicitation of a full preference order. These additional observations per choice set offer three potential benefits. Firstly, a smaller sample size is required to obtain similar efficiency, that is, similar standard errors, which might be useful if there are budget constraints, or if the research population is small (Hausman & Ruud, 1987; Lancsar et al., 2013; Palma, 2017; Vermeulen et al., 2011). Otherwise, one can choose to work with the same sample size, but present fewer choice sets to each respondent to decrease respondent burden and still have similar efficiency. Lastly, it facilitates more advanced analyses like the estimation of individual-level models (Hess & Rose, 2009; Lancsar & Louviere, 2008; Louviere, Street, et al., 2008). Especially in health, these potential benefits have encouraged researchers to move beyond single-best DCE.

Full preference orders that go beyond single-best DCE can be elicited by (1) ranking all alternatives simultaneously, (2) repeatedly choosing the best and the worst alternatives, or (3) repeatedly choosing the best alternatives. Ranking has been used in DCEs for many years (Beggs et al., 1981; Chapman & Staelin, 1982; Punj & Staelin, 1978). It provides a full preference order, and the sequence in which alternatives are ranked can be specified by the researcher or left up to the respondent. However, it has been argued to be burdensome and the stability of the results have been found to differ per rank (Ben-Akiva et al., 1991; Bradley & Daly, 1994; Fok et al., 2012; Hausman & Ruud, 1987). Best-worst DCE (BWDCE) is another method to elicit partial or full preference orders that has been proposed to overcome cognitive burden (Marley & Louviere, 2005). In BWDCE, respondents are asked to choose the best and worst alternative from a set of at least three alternatives.[1] This can be sequential (i.e., first choosing the best alternative, then worst and repeat this for the remaining alternatives until all alternatives are eliminated) or simultaneous (i.e., choosing the best and worst alternative at the same time and repeat this until a full preference order is reached). Regardless of the way in which best and worst choices are elicited, there are studies that question whether positive (i.e., best) choices and negative (i.e., worst) choices have the same underlying mental processes (Delle Site et al., 2019; Dyachenko et al., 2014; Hawkins et al., 2014, 2019). As outlined in Dyachenko et al. (2014), in addition to the decision sequence affecting preferences if best and worst choices are elicited sequentially, the mere framing of best and worst choices may lead to two different mental processes. To overcome this, the method best-best DCE (BBDCE) was introduced relatively recently to avoid the need to swap between positive and negative mental processes (Ghijben et al., 2014). This third method to move beyond traditional single-best DCE is characterized by eliciting single-best choice, second-best choice from the remaining alternatives and so on until a full preference order is obtained.

If the objective is to obtain more preference information per choice set than from a traditional single-best DCE, key, to date unanswered, questions are (1) which of BWDCE, BBDCE and ranking is most appropriate to do so and (2) what are potential downsides if using one of these elicitation methods.

We identified 15 studies that empirically compared elicitation methods that move beyond single-best DCE. These can be arranged into three categories, namely studies that explored differences in choices (a) within a preference elicitation method (e.g., differences between best and worst choices in BWDCE); (b) between preference elicitation methods (e.g., differences between BWDCE and ranking); or (c) both within and between methods. None of these studies provide a head-to-head comparison of choices within and between BWDCE, BBDCE and ranking DCE, that is, (c) above with all three methods. Instead, they were restricted to a comparison of (usually subsets of) best-worst, ranking, and traditional single-best DCE and only one included BBDCE. Most were performed outside health. We provide an overview of the identified studies below and their characteristics in Table A1 with their results discussed in more detail in Supporting Information S1.

Six studies (Ben-Akiva et al., 1991; Delle Site et al., 2019; Hausman & Ruud, 1987; Hawkins et al., 2019; Lancsar et al., 2013; Scarpa et al., 2011) were of type (a), meaning preferences were elicited using the same method but first, second and/or follow-up choices were analyzed separately. These studies compared choices within a preference elicitation method, but they did not compare between different elicitation methods. Only Lancsar et al. (2013) studied choices in the health domain. They generally found differences in preferences, scale, and/or efficiency between choices. Another six studies (Akaichi et al., 2013; Caparrós et al., 2008; Chang et al., 2009; Petrolia et al., 2018; Xie et al., 2014; Yangui et al., 2019) experimentally compared a subset of the methods that move beyond DCE or compared one or two of these methods to traditional single-best DCE, that

is, (b) above. These studies compared different elicitation methods but did not compare choices within a preference elicitation method. They found differences in preferences, respondent burden, predictive ability, and/or scale. Only Xie et al. (2014) studied choices in the health domain.

Hawkins et al. (2014), Krucien et al. (2019), and Giergiczny et al. (2017) studied differences both within and between methods, that is, (c) above, but found conflicting results regarding similarity of preferences and none compared ranking, BBDCE and BWDCE. None compared all three of ranking, BBDCE and BWDCE. Krucien et al. (2019) was the only study that compared choices within and between methods in the health context. Hawkins et al. (2014) compared three types of elicitation methods: best choice, worst choice, and best-worst choices. They found that the first choices were similar, regardless of the presence of a second choice. Also, best and worst choices were inversely related, based on which the authors conclude that they are based on the same underlying information. Krucien et al. (2019) compared the same methods as Hawkins et al. (2014) but found that preferences differed between choices, even beyond scale. Choice consistency was higher in best choices. Giergiczny et al. (2017) compared best-best and best-worst elicitation methods and single-best, best-best and best-worst model specifications. They found in both elicitation methods that willingness to pay estimates from models using all choices were significantly different from estimates in models using only first choices, even if correcting for scale. However, preferences and scale parameters were similar between elicitation methods if the data were modeled using the same (recoded) exploded logit specification.

While all these studies provide interesting and varied insights into the various elicitation methods and their first, second and/or follow-up choices, none provide a comparison within and between methods in best-worst, best-best and ranking DCEs. As such, the primary objective of this study was to undertake a head-to-head comparison of choices within and between methods in BWDCE, BBDCE, and ranking DCEs to inform the decision on which preference method to use if wanting to move beyond traditional single-best DCE. We focus on the health domain, namely by eliciting preferences for Australian obesity reduction policies. To provide a practical overview of the advantages and disadvantages of all three methods, we compare them using six criteria. Four criteria are defined using choice modeling estimates, namely: (1) trade-off consistency; (2) choice consistency; (3) scale dynamics, interpreted as learning or fatigue effects; and (4) statistical efficiency. The remaining two criteria are based on respondent-reported data, namely (5) difficulty and (6) respondent stated preference between the three preference elicitation approaches. Each criterion is defined in the methods section, along with an explanation of the study format and the details of the analyses. Results are presented per assessment criterion, followed by a discussion of the findings.

2 | METHODS

2.1 | Study format

To enable a direct comparison between BWDCE, BBDCE and ranking DCE, the study consisted of three treatment arms that each represented one of the methods. Importantly, the experiments per arm were identical (same choice sets, versions etc.) except that the elicitation task, that is, the choice questions asked, differed. Respondents were randomized to one of three arms. The survey involved four parts. Firstly, respondents answered screening questions that enabled quota sampling. Secondly, respondents were presented with 16 DCE choice sets (see "Attributes, levels and data collection"). In the BWDCE and BBDCE arms respondents faced two choices per choice set. In the first choice they were asked to choose their most preferred alternative. This chosen alternative was then removed from the choice set. Respondents were subsequently asked to choose their worst or best alternative from the remaining alternatives, respectively in best-worst and best-best. The resulting choices implied a complete preference order over the three alternatives per choice set in each of these elicitation methods. In the ranking arm, respondents were asked to provide an explicit preference order by placing a one under the alternative they most preferred, a two under the alternative they next most preferred and a three under the alternative they preferred least. All alternatives were visible until the respondent provided the full preference order and moved to the next choice set. Examples of the choice tasks can be found in Supporting Information S2. Thirdly, respondents answered questions about the DCE. In particular, they stated the perceived difficulty of the experienced elicitation method and their preference between each of the three elicitation methods: one of which they had experienced, the other two were described to them. Lastly, respondents answered socio-demographic, health-related and attitudinal questions relating to obesity.

2.2 | Attributes, levels and data collection

The study concerned taxpayer preferences for healthcare policies to reduce obesity in Australia. For more information about the attributes, levels, experimental design, and the online survey design, please see Lancsar et al. (2021) which explores preferences

based on single-best data. Attributes in the study were policy type, effectiveness in reducing obesity rates, and cost in terms of higher taxes. The attributes and their levels are described in Table 1.

A total of 256 choice sets were blocked into 16 versions of 16 choice sets. As mentioned earlier, respondents were randomized to one of three arms. Within each arm respondents were randomized to one of the 16 versions of 16 choice sets. Each choice set consisted of three alternatives. Two alternatives were generic and designed to vary per choice set (Policy A and Policy B), the other alternative was a constant status quo that reflected the current situation. This current situation was described as no additional policy interventions, no change to the projected obesity rate (i.e., 32% would be obese in 2020) and no additional cost. The survey was administered to a sample of taxpayers from an online panel, representative of the Australian population of taxpayers in age and gender. The sample size was chosen to allow estimation of reliable models in each arm, while considering the number of attributes, levels, respondents per version and the parameters to be estimated (Lancsar & Louviere, 2008). Respondents were asked to suppose the Australian government was considering introducing new policies to reduce obesity and is interested in taxpayer preferences.

2.3 | Statistical analysis

The discrete choice data in this study are modeled using random utility theory. To capture all information contained in a full preference order, we used rank-ordered (also called exploded) mixed logit models (Train, 2003). We first elaborate on this modeling approach, followed by a description of the six assessment criteria used to compare the preference elicitation methods. Lastly, we describe the four different stages of analysis.

Following random utility theory, each respondent i makes choices concerning the alternatives in a choice set t based on the latent utility of each alternative j, given by U_{ijt}. The utility is comprised of a systematic component $X'_{ijt}\beta_i$ and a random component ε_{ijt} :

$$U_{ijt} = X'_{ijt}\beta_i + \varepsilon_{ijt}, \tag{1}$$

where β_i is an individual-specific vector describing the preference weights of the attributes and X_{ijt} represents the attribute levels for individual i, alternative j in choice set t. Following Swait and Louviere (1993), one can extend (1) to allow for variation in

TABLE 1 Attributes and levels

Attribute	Levels
Policy type	Nutritional information labeling using traffic light symbols
	National mass media campaign to encourage healthy lifestyle choices
	Ban unhealthy food and drink advertising to children
	Improve nutritional quality of foods sold in public institutions
	Funding for physical activity infrastructure and outdoor spaces
	Tax sugar-sweetened beverages
	Payment incentive for the obese to increase physical activity
	Pre-paid cards for healthy foods in supermarkets (reference level)
Impact on obesity rates in 2020 (effectiveness)	32% will be obese in 2020 (no change to the projected obesity rate)
	31% will be obese in 2020 (moderate reduction in the projected obesity rate)
	29% will be obese in 2020 (large reduction in the projected obesity rate)
	28% will be obese in 2020 (very large reduction in the projected obesity rate)
Additional cost to you per year, paid as an increase in income taxes by (cost)	$12 per year ($1 per month)
	$120 per year ($10 per month)
	$240 per year ($20 per month)
	$480 per year ($40 per month)

choice consistency by including a scale parameter λ. This scale parameter can differ per elicitation method and/or choice and scales the systematic component of utility[2]:

$$U_{ijt} = \lambda_s \cdot \left(X'_{ijt} \beta_i \right) + \varepsilon_{ijt}. \tag{2}$$

To identify the scale effects, we impose $\lambda_1 = 1$ for the reference scenario, where a scenario refers to the combination of the elicitation method and the choice within that method. If mean preference weights over all individuals in the sample (β) are stable across elicitation methods and/or choices, beyond scale, then one would obtain mean preference weights specific to a scenario s: $\beta_s = \lambda_s \beta_1$ (Louviere, Islam, et al., 2008). However, when not only the relative importance of the error term is different, but also the trade-offs as characterized by the mean preference weights β, then this approach can no longer be used to capture scale differences. Instead, we build on the ideas of the McKelvey and Zavoina (1975) pseudo R-squared, given by $R^2 = \frac{Var(X'\beta)}{Var(X'\beta) + \sigma_\varepsilon^2}$, which captures the relative importance of the systematic part in choice models using the variance of $X'\beta$. Using the variation in the systematic utility component within choice sets, we generalize the definition of the utility scale λ as follows:

$$\lambda_s = \sqrt{\frac{\sum_{t=1}^{S} \sum_{j \neq k} \left(\left(X_{t,j} - X_{t,k} \right) \beta_s \right)^2}{\sum_{t=1}^{S} \sum_{j \neq k} \left(\left(X_{t,j} - X_{t,k} \right) \beta_1 \right)^2}}. \tag{3}$$

Here β_s and β_1 represent the mean preference weights in the elicitation method and/or choice under evaluation and the reference scenario, respectively. Furthermore, $X_{t,j}$ and $X_{t,k}$ represent the vectors of attributes of alternatives j and k in choice set t, respectively. This definition relies on the notion that choices are more consistent when the differences between the systematic utility components of the alternatives in the choice set become larger. Note that this indeed generalizes the usual definition of the utility scale when $\beta_s = \lambda_s \beta_1$.

To also capture variation in choice consistency over the course of the choice sets, the scale parameter λ_s is defined using a combination of γ_s, the choice consistency of the first choice set (i.e., $t = 1$), and δ_{st}, the linearly specified[3] scale dynamics over the course of the choice sets, relative to $t = 1$ (i.e., from task $t = 2$ to $t = 16$). To ensure a positive value for λ_s, a log-linear specification is used:

$$\log \left(\lambda_{st} \right) = \gamma_s + (t - 1) \cdot \delta_{st}, \tag{4}$$

where, for identification purposes, γ_s is set to zero if β can differ between elicitation methods and/or choices.

By moving beyond single-best DCE, more information is elicited from one respondent or one choice set. To account for this multitude of choices per choice set, we used a rank-ordered logit (also called exploded logit) model as first introduced by Punj and Staelin (1978). Furthermore, to also capture heterogeneity between individuals, and hence to estimate the individual-specific preference weights β_i as specified in Equation (1), rank-ordered mixed logit models were used (Train, 2003). Note that as explained in study format, in BWDCE and BBDCE respondents faced two sequential choices per choice set. In ranking, respondents were asked to provide an explicit one-off preference order. As each choice set had three alternatives, all three methods resulted in a complete preference ordering. This allows analysis using the same exploded logit version of the likelihood. By keeping everything constant across elicitation methods, including the analyses, observed differences between elicitation methods and/or choices can be directly attributed to the way in which the preferences were elicited.

2.3.1 | Criteria

Six assessment criteria are used to compare the methods best-worst, best-best and ranking. The first four criteria are based on the choice modeling approach described above; they are (1) *trade-off consistency*, (2) *choice consistency*, (3) *scale dynamics* and (4) and *statistical efficiency*. The last two criteria are based on respondent-reported variables: (5) *difficulty* and (6) *preference*. Now we provide further explanation of the criteria; a summary of the criteria and their operationalization is provided in Table 2. Choice modeling analyses were performed in Julia version 1.6.1. Other statistical analyses were performed in R version 4.1.2.

TABLE 2 Summary assessment criteria

Criterion	Operationalization	Values and interpretation	Preferred direction
Trade-off consistency	Marginal rate of substitution (MRS); poolability	Zero to infinity	No preferred direction for MRS value; similarity and poolability are preferred
Choice consistency	Relative scale differences in systematic utility (λ)	Compared to ref. scenario: 0–1: Choices less consistent 1: Choices equally consistent >1: Choices more consistent	Higher values are preferred
	Log of the scale of the first choice set (γ)	Compared to ref. Scenario: <0: Choices less consistent 0: Choices equally consistent >0: Choices more consistent	Higher values (less negative) are preferred
Scale dynamics	Change in error scales over choice sets (δ)	<0: Fatigue 0: Neutral >0: Learning	Close to zero is preferred over positive or negative values
Statistical efficiency	Precision of MRS ($\frac{1}{\sqrt{N}\cdot\sigma}$)	Zero to infinity	Higher values are preferred
Difficulty	Respondent-reported (easy, neutral, difficult)	0%–100% easy 0%–100% neutral 0%–100% difficult	Higher values are preferred if easy, lower values are preferred if difficult
Preference	Respondent-reported (yes, no)	0%–50%: Experienced method is preferred 50%–100%: Described method is preferred	Lower values are preferred if 0%–50%, higher values are preferred if 50%–100%

2.3.2 | Trade-off consistency

We assess trade-off consistency by comparing the marginal rates of substitution (MRS) of attributes between elicitation methods and/or choices. This MRS between two attributes X_k and X_l can be computed as:

$$\text{MRS} = \frac{\partial U / \partial X_k}{\partial U / \partial X_l}. \tag{5}$$

Note that the MRS simplifies to the ratio of the preference weights β_k and β_l if the utility function is specified to be linear in parameters. In this study, the ratio of the mean preference weights is studied and standard errors of the MRS are approximated using the parametric bootstrap (Krinsky & Robb, 1986; Park et al., 1991), that is, by randomly sampling for each MRS from the corresponding multivariate normal distribution with mean preference weights and covariance matrices of the random parameters β_k and β_l and calculating the ratio using 10,000 replications. We focus on the MRS between cost and effectiveness for its ease of interpretation, other MRS are also reported. From a methodological perspective, there is no preferred value of the MRS. However, similar values between elicitation methods and/or choices are preferred, as these indicate that preferences do not vary between elicitation methods or choices.

Furthermore, in line with Swait and Louviere (1993) consistency in preferences across elicitation methods and/or choices is assessed by testing whether parameters in the various elicitation methods and/or choices are similar (i.e., can be pooled) while allowing for differences in scale per elicitation method, per choice or both. Poolability of parameters across elicitation methods and/or choices is formally tested using likelihood-ratio tests. Again, similarity of estimates and hence poolablity of preference weights is preferred.

2.3.3 | Choice consistency

The second criterion, choice consistency, assesses whether differences between the systematic utility components of the alternatives in the choice set are larger, relative to the scale of the error term, resulting in higher choice consistency. Higher choice

consistency means lower variability in choices. Firstly, if β_s is estimated separately for the elicitation methods and/or choices, we capture choice consistency using λ_s, the scale of the systematic utility component defined in Equation (3). A value below one indicates that choices are less consistent than the reference scenario, a value above one indicates choices are more consistent than the reference scenario. Higher choice consistency is preferred. Secondly, if β_s is restricted to be the same (up to scale) for the elicitation methods and/or choices (i.e., aggregate-level analysis), we capture choice consistency using γ_s, the choice consistency of the first choice set as defined in the log-linear specification in Equation (4). Again, higher choice consistency is preferred. Values of γ_s below zero, that is, log (1), indicate lower choice consistency, values above zero indicate higher choice consistency. It should be noted that this measure only indicates how (in)consistent choices are for a given number of alternatives, and hence cannot be used to make a comparison between choice sets with different numbers of alternatives.

2.3.4 | Scale dynamics

Thirdly, scale dynamics indicates choice consistency over the course of the choice sets. Increasing unexplained variance and hence lower choice consistency over the choice sets could indicate fatigue (Bradley & Daly, 1994). Vice versa, if choice consistency is increasing, there could be learning effects. In this study, scale dynamics is quantified with δ_{st}, defined in Equation (4), which is related to the scale of the systematic component of utility.[2] A negative value of δ_{st} indicates that scale is decreasing, and that unexplained variance is increasing over the choice sets, and hence we interpret it as fatigue. Similarly, we interpret positive values as evidence of learning. We consider values close to zero to be preferable. Learning and fatigue both indicate that choices might be too difficult at the beginning or end of the series of choice sets, respectively.

2.3.5 | Statistical efficiency

The last choice modeling criterion is statistical efficiency of the parameter estimates. As mentioned, if more information is elicited from one respondent or one choice set, a smaller sample size or fewer choice sets per respondent is required to obtain similar efficiency (i.e., similar precision of estimates). Our measure of efficiency uses the standard error of the MRS as in Equation (5). We use as measure of efficiency of a method the obtained standard error of the MRS, corrected for the impact of sample size, as this is informative about the information content and hence efficiency of a single respondent. Specifically, our efficiency measure is defined as:

$$\text{Efficiency} = \frac{1}{\sqrt{N_s} \cdot \sigma_s}. \tag{6}$$

Here N_s is the sample size used in the elicitation method and/or choice s under evaluation and σ_s is the standard error of the estimated MRS on the sample of N respondents in that particular elicitation method and/or choice. This measure is tightly connected to the efficiency analysis and sample size calculations in Rose and Bliemer (2013), with the required sample size for a certain precision of the MRS being proportional to this efficiency measure.

2.3.6 | Difficulty

The first criterion based on respondent-reported data is difficulty. All respondents were asked to report the difficulty of the (sub) choices they experienced by picking which description best reflects their experience: easy, difficult, or neither. These results are compared between elicitation methods and first and second choices using chi-squared tests. The exact phrasing of the question is reported in Table A2.

2.3.7 | Preference

The last criterion, *preference,* is also measured using respondent-reported data. To assess this, respondents in each arm (i.e., in each method) were asked whether they would have preferred another (described) method than the one they experienced. Results are compared descriptively between elicitation methods. The phrasing of the questions is reported in Table A2.

2.3.8 | Four-stage analysis

To evaluate the preference elicitation methods, the criteria are compared in four stages. Firstly, trade-off consistency, choice consistency, scale dynamics and efficiency are evaluated on a disaggregate level, meaning that parameters are separately estimated for first and second choices in all three elicitation methods. In the second stage of analysis, first and second choices are aggregated within the elicitation method. These aggregate-level analyses allow a direct comparison between the three methods best-worst, best-best and ranking in terms of trade-off consistency, choice consistency, scale dynamics and efficiency. Thirdly, to assess first and second choices regardless of the elicitation method, parameters are estimated by aggregating per choice across the three methods. This allows comparison between first and second choices in terms of trade-off consistency, choice consistency, scale dynamics and efficiency. Lastly, descriptive analyses are performed to assess the criteria difficulty and preference. A summary of the assessment criteria can be found in Table 2.

3 | RESULTS

No statistically significant differences between arms in age, gender, employment status, general health, education and survey response time were identified using Bonferroni adjusted tests. Respondent characteristics per arm can be found in Table A3 in the Appendix.

Table 3 provides a concise overview of the results across methods and across criteria. The tick marks indicate which elicitation method or choice performed best on a particular criterion, multiple tick marks per criterion implies that methods or choices performed equally well, a hyphen indicates that we could not draw a conclusion based on the available data. We further elaborate on these findings below and in the discussion.

Before continuing to the results, we note the following. In the disaggregate analyses, the parameters were estimated for every elicitation method and choice separately. Preferences were not the same across the six choices, not even up to scale implying strictly speaking one should not aggregate preferences within or between methods. However, following the aim of this study to help researchers decide which method to use if moving beyond traditional single-best DCEs, we do present how well the methods performed on every criterion if one would pool choices within a method (i.e., aggregate per elicitation method). Additionally, we also want to inform the decision on whether to even move beyond single-best DCE by presenting how well first choices performed compared to second choices (i.e., aggregate per choice). The remainder of this section describes the results per stage of analysis and per criterion. For brevity, only a subset of the estimation results is presented; the full set of estimation results are found in Supporting Information S3. Tables OSM 1–6.

3.1 | Disaggregate analyses per method and choice

Table 4 presents the results of the model on a disaggregate level, meaning that the model parameters were estimated for every elicitation method and choice separately.

3.1.1 | Trade-off consistency

In the first choices of the three methods, the MRS between cost and effectiveness was lower ($5.44/%–$5.77/% increased effectiveness) than in the second choices ($6.40/%–$7.71/%). A similar pattern was found in the MRS for the majority of the

TABLE 3 Summary results per criterion and per stage of analysis

Criterion	Disaggregate per method and choice						Aggr. per elicitation method			Aggr. per choice	
	BW1	BW2	BB1	BB2	R1	R2	BW	BB	Rank	1st	2nd
Trade-off consistency	-	-	-	-	-	-	-	-	-	-	-
Choice consistency	√						√	√		√	
Scale dynamics	√		√	√		√			√	√	
Statistical efficiency	√			√	√			√		√	
Difficulty					√				√	√	
Preference			√	√			√		√	-	-

Abbreviations: BB, best-best; BW, best-worst.

TABLE 4 Disaggregate results

Variables	BW1 Est.	BW1 SE	BW2 Est.	BW2 SE	BB1 Est.	BB1 SE	BB2 Est.	BB2 SE	Rank1 Est.	Rank1 SE	Rank2 Est.	Rank2 SE
MRS ($\frac{\partial U/\partial \text{eff}}{\partial U/\partial X^{\text{cost}}}$)	$5.77/%	0.469	$7.71/%	0.874	$5.72/%	0.593	$6.69/%	0.743	$5.44/%	0.484	$6.40/%	0.579
Choice consistency (λ)	REF		0.393		0.730		0.731		0.691		0.610	
Scale dynamics (δ)	0.001	0.005	0.033	0.006	0.008	0.005	0.004	0.006	0.010	0.004	−0.001	0.006
Efficiency ($\frac{1}{\sqrt{N}\cdot\sigma}$)	0.067		0.036		0.055		0.044		0.065		0.054	
N parameters	2 × 36				2 × 36				2 × 36			
N respondents	997				926				1012			
Log-likelihood	−20552.3				−17928.9				−19778.7			

Abbreviations: BB, best-best; BW, best-worst.

policies, as shown in Supporting Information S3: Table OSM 2. Furthermore, based on the likelihood ratio (LR) tests for poolability, preferences were not the same across all six sub choices (test-statistic = 2262.8, df = 175, $p < 0.001$), not even up to scale (test-statistic = 1827.8, df = 170, $p < 0.001$).

3.1.2 | Choice consistency

Choice consistency, quantified with λ as specified in Equation (3), was below one in all choices in all elicitation methods, indicating lower choice consistency than the reference scenario (the first choice of best-worst). In best-best and ranking, first and second choices within a method were relatively similar. Overall, responses were most consistent in the first choices of best-worst and least consistent in the second choices of best-worst.

3.1.3 | Scale dynamics

Besides the second choices of ranking, scale dynamics (i.e., scale over the course of the choice sets) were positive in all choices and elicitation methods, but statistically significantly different to zero only for the second best-worst choice and first ranking choice. This indicates that in these cases, unexplained variance was decreasing over the choice sets, which could be interpreted as learning.

3.1.4 | Statistical efficiency

Statistical efficiency as specified in Equation (6), connected with the inverse of the standard errors of the MRS between cost and effectiveness, was higher in the first choices as compared to the second choices. Since the first choice is a choice out of all three, and the second choice is a choice out of two statistical efficiency is expected to be higher in the first than the second choices (Vanniyasingam et al., 2016). In the first choices efficiency was highest in best-worst, in the second choices it was highest in ranking. However, most efficiency parameters for the different policies, as shown in Supporting Information S3: Table OSM 2, show that efficiency of the first choices was highest in ranking. For the second choices, it was generally highest in best-best and lowest in best-worst.

3.2 | Aggregate analyses per elicitation method

A summary of the results in which preferences were estimated for every elicitation method separately, but preferences of the choices were aggregated, that is, assumed identical, within the method can be found in Table 5. Choice consistency and scale dynamics were allowed to be different for first and second choices, as pooling tests showed that these parameters could not be aggregated.[4]

TABLE 5 Aggregate results per elicitation method

Variables	BW Est.	SE	BB Est.	SE	Rank Est.	SE
MRS ($\frac{\partial U/\partial \text{eff}}{\partial U/\partial X^{\text{cost}}}$)	$6.10/%	0.687	$6.64/%	0.494	$6.43/%	0.511
Choice consistency (λ) 1st	REF		0.850		0.828	
Choice consistency (γ) 2nd	−1.024	0.076	−0.172	0.000	−0.247	0.060
Scale dynamics (δ) 1st	−0.004	0.005	0.006	0.005	0.009	0.005
Scale dynamics (δ) 2nd	0.033	0.006	0.014	0.007	0.005	0.006
Efficiency ($\frac{1}{\sqrt{N \cdot \sigma}}$)	0.046		0.066		0.061	
N parameters	38		38		38	
N respondents	997		926		1012	
Log-likelihood	−20982.5		−18182.9		−19872.2	
Pooling test for combining two choices within method	860.4, df = 34, $p < 0.001$		508.0, df = 34, $p < 0.001$		187.0, df = 34, $p < 0.001$	

Abbreviations: BB, best-best; BW, best-worst.

3.2.1 | Trade-off consistency

The MRS between cost and effectiveness was $6.10/% increase in effectiveness in best-worst, $6.64/% in best-best and $6.43/% in ranking. For the policies, the MRS was generally lowest for best-best and highest for best-worst, as shown in Supporting Information S3: Table OSM 4. However, for all methods the model that pooled choices within the method was rejected in favor of the model with separate first and second choices (LR test statistics 860.4, 508.0, and 187.0 for BW, BB and ranking, respectively, all df = 34, $p < 0.001$). These results of the pooling test show that even if allowing differences in choice consistency and scale dynamics between choices, preferences of the first and second choices could not be pooled in any of the methods, and hence first and second choices should be analyzed separately. The results below should therefore be interpreted with caution. They show how the elicitation methods would perform if disregarding the fact that preferences from first and second choices could not be pooled for any of the methods.

3.2.2 | Choice consistency

Like preferences, choice consistency could not be pooled between the first and second choices. As such, choice consistency was measured using two different estimates for choice consistency in these analyses. Firstly, choice consistency (λ) as specified in Equation (3) was below one in best-best and ranking, indicating less consistent responses in the first choices of those elicitation methods than in the reference scenario best-worst. Secondly, a separate estimate of choice consistency (γ) as specified in Equation (4) was estimated for the second choices in each elicitation method, as compared to the choice consistency of that method's first choices. Here, the negative parameters indicate that second choices were less consistent than first choices. In best-best, second choices were closest to the first choices in terms of consistency.

3.2.3 | Scale dynamics

Most scale dynamics parameters were positive, suggesting that in most elicitation methods respondents experienced learning. However, it was only significant for the second choices of best-worst and best-best. In ranking, none of the parameters were statistically different from zero suggesting the least learning over the course of the choice sets in this method.

3.2.4 | Statistical efficiency

Statistical efficiency of the preference estimates as computed using the standard errors of the MRS between cost and effectiveness was highest for best-best, then ranking and least for best-worst. A similar pattern was observed in all policy parameters as shown in Supporting Information S3: Table OSM 4.

3.3 | Aggregate analyses per choice

Table 6 presents a summary of the results of the model in which preferences were estimated for first and second choice separately but pooled between elicitation methods. Like before, choice consistency and scale dynamics were allowed to differ between methods, as pooling tests showed that these parameters could not be aggregated (results not reported).

3.3.1 | Trade-off consistency

The MRS between cost and effectiveness was $5.61/% effectiveness in the first choices and $7.59/% in the second choices. As shown in Supporting Information S3: Table OSM 6, the MRS was generally lower in first than in second choices. However, neither the first nor second choices could be pooled between elicitation methods (LR test statistics 132.2, and 224.0 for first and second choices, respectively, both df = 68, $p < 0.001$). This shows that even after allowing for differences in choice consistency and scale dynamics between methods, preferences could not be pooled between methods, and hence first and second choices should be analyzed separately. Based on this finding, again, the results below should be interpreted with caution.

3.3.2 | Choice consistency

As in the aggregate analysis, we report two different types of choice consistency: choice consistency (λ) as specified in Equation (3) if preferences are estimated separately per choice across methods, and choice consistency (γ) as specified in Equation (4) if preferences are assessed in aggregate. Firstly, in the second choices, choice consistency (λ) was below one, indicating less consistent choices than in the reference scenario (i.e., first choices of best-worst). Secondly, separate estimates of choice consistency (γ) were estimated for the choices of the other elicitation methods, as compared to the choice consistency of best-worst. Within first choices, the positive parameter for ranking indicates that first choices in ranking were more consistent than first choices of best-worst. Within the second choices, responses in best-best and ranking were more consistent than the second choices in best-worst. Furthermore, the estimates of choice consistency (γ) in the first choices were closer to zero than in the second choices, indicating more similar choice consistency of responses in first than second choices.

3.3.3 | Scale dynamics

Like observed when aggregating preferences over the elicitation methods, most scale dynamics parameters were positive, suggesting that most respondents experienced learning. Besides in ranking, scale parameters of first choices were closer to zero than second choices meaning that first choices are preferred. Scale dynamics were especially large in the second choice of best-worst.

TABLE 6 Aggregate results per choice

Variables	1st Est.	SE	2nd Est.	SE
MRS ($\frac{\partial U/\partial \text{eff}}{\partial U/\partial X^{\text{cost}}}$)	$5.61/%	0.370	$7.59/%	0.450
Choice consistency (λ) BW	REF	NA	0.510	NA
Choice consistency (γ) BB	−0.066	0.065	0.340	0.089
Choice consistency (γ) rank	0.011	0.067	0.295	0.086
Scale dynamics (δ) BW	0.011	0.005	0.037	0.007
Scale dynamics (δ) BB	0.010	0.005	0.018	0.007
Scale dynamics (δ) rank	0.009	0.005	0.007	0.006
Efficiency ($\frac{1}{\sqrt{N \cdot \sigma}}$)	0.086		0.073	
N parameters	40		40	
N respondents	2935		2935	
Log-likelihood	−33499.4		−25113.8	
Pooling test for combining three choices between methods	132.2, df = 68, $p < 0.001$		224.0, df = 68, $p < 0.001$	

3.3.4 | Statistical efficiency

The MRS between cost and effectiveness showed highest statistical efficiency for the first choices. This also holds for all policy parameters, as shown in Supporting Information S3: Table OSM 6.

3.4 | Descriptive analysis

Descriptive results provided additional insights into respondent stated perceived difficulty and preference of respondents regarding which of best-worst, best-best and ranking DCEs they thought were "best" and are presented in Table 7.

3.4.1 | Difficulty

Difficulty was found to vary between choices and elicitation methods. The majority of respondents who were randomized to the ranking task perceived the ranking task to be easy (69%), higher than in best-worst (59%–63% depending on first or second choice) and best-best (49%–62%). Respondents who were randomized to the best-worst and best-best tasks both more often perceived the first choice to be easy (63% and 62% respectively) than the second choice (59% and 49% respectively). The second choice in best-best was most often reported to be difficult (14%). The differences in perceived difficulty were statistically significant for nearly all choices ($p < 0.001$), no statistically significant differences were found between the first choice of best-worst and best-best ($p = 0.77$).

3.4.2 | Preference

At the end of the survey, the two elicitation methods respondents did not experience were described to them and they were asked whether they would prefer each over the one they just completed (See Table A2 for phrasing of questions). Some respondents would have preferred another method than the one they experienced (27%–41% depending on the experienced and described method). In the ranking arm, the lowest number of respondents would have preferred another method over the one they experienced. In the best-best arm, 38% would have preferred best-worst and 38% would have preferred ranking. In the best-worst arm, the highest number of respondents would have preferred another method than best-worst. These results suggest ranking is the most preferred method, followed by best-best, and then best-worst.

4 | DISCUSSION

This study undertook a head-to-head comparison of BWDCE, BBDCE and ranking DCEs between and within methods, to inform the decision on which preference method to use when considering moving beyond traditional single-best DCE. Eliciting additional observations per choice set can translate into needing a smaller sample size, reducing respondent burden with fewer

TABLE 7 Difficulty and preference per elicitation method—descriptive results

	Best-worst N = 997		Best-best N = 926		Ranking N = 1012
How did you find the task of…[a]	Best	Worst	Best	2nd best	Ranking
Easy	63%	59%	62%	49%	69%
Neither	32%	32%	33%	37%	26%
Difficult	5%	9%	5%	14%	4%
% Who would have preferred[b]					
Best-best	38%				36%
Best-worst			38%		27%
Ranking	41%		38%		

[a]Statistically significant differences between all choices ($p < 0.01$) besides difference between first choice in best-worst and best-best ($p = 0.77$). See Table A2 for exact phrasing of question.

[b]Percentages don't add up to 100% because questions were not mutually exclusive. See Table A2 for phrasing of questions.

choice sets for each respondent and/or enabling more advanced analyses. We compared the preference elicitation methods using six criteria, namely: (1) trade-off consistency, (2) choice consistency, (3) scale dynamics, also known as learning or fatigue effects, (4) statistical efficiency, (5) stated difficulty and (6) stated preference. These were tested in a health-based study that elicited preferences for obesity reduction policies in Australia.

Most importantly, we found that preferences were not the same across the methods and their sub-choices, not even up to scale, based on poolability testing. This implies that in this study, the first to compare all three methods, we should not aggregate preferences within or between methods. However, for some researchers the benefits of moving beyond single-best DCE weigh strongly. To inform their decision on which preference method to use if moving beyond traditional single-best DCEs, we analyzed the methods' performance on every criterion by means of aggregate analyses per elicitation method. Ranking performed best in terms of scale dynamics, respondent-reported difficulty, and respondent-reported preference. Best-worst performed best in terms of choice consistency, while statistical efficiency of the preference estimates was highest for best-best.

In the disaggregate analyses, where we separately analyze first and second choices, the first choices of best-worst performed well in terms of choice consistency and scale dynamics and statistical efficiency, but this method's second choices performed worst on these criteria. First choices from ranking and best-best performed less well than those of best-worst on these criteria but their second choices performed better than those of best-worst. Of these two methods, ranking performed better on efficiency and best-best on choice consistency. The low efficiency of each individual choice in best -best is in contrast with its high performance when pooled. This might be due to the two choices being more similar—although not poolable—for best-best. The results of this study can also be used to inform the decision on whether to even move beyond single-best DCE by comparing first choices to second choices (i.e., aggregate per choice). We found that first choices outperformed second choices on all criteria, but that neither the first nor second choices could be pooled across elicitation methods.

With preferences from second choices and especially worst choices reflecting different trade-offs, combining them with first choices will make the overall results less accurate and potentially invalid. Furthermore, with preferences from first choices differing between elicitation methods, we suggest using a traditional single-best DCE would generally be most appropriate. However, if one considers moving beyond first choice, for example, due to a tight budget or if the research population is small, based on our empirical results we recommend best-best or ranking. Our results provide an empirical basis for the decision to move beyond single-best DCE and choosing an elicitation method to do so, however, the actual decision will always depend on the criteria at hand.

These differences in elicitation methods and choices are in line with the empirical work that found that preferences were different, depending on method or choice (Ben-Akiva et al., 1991; Delle Site et al., 2019; Hausman & Ruud, 1987). In particular, the large differences between best and worst choices are in line with literature that states that attributes and alternatives are weighted differently depending on whether the choice is to accept or reject (Dhar & Wertenbroch, 2000; Laran & Wilcox, 2011; Meloy & Russo, 2004; Shafir, 1993) as this likely shifts a person's mindset from a promotion to a prevention focus (Brockner & Higgins, 2001; Higgins et al., 1997). Furthermore, like Giergiczny et al. (2017) and Krucien et al. (2019), we found that first choices are different from second choices, which raises doubts regarding whether to move beyond traditional single-best preference elicitation methods. If the aim is to measure what people want most, then this is measured well with the first choice (de Bekker-Grob et al., 2019, 2020).

This study is the first to provide a head-to-head comparison of the three elicitation methods, including the relatively new method BBDCE that had been hypothesized to overcome some of these difficulties by using the same mindset for each decision (Ghijben et al., 2014). This extensive comparison between various elicitation methods has not yet been made in health nor in the literature more broadly. There are, however, some aspects of the study worth noting to allow careful interpretation of the findings. Firstly, from a methodological point of view, we note that respondents had to choose between three alternatives with two sequential choices in BWDCE and BBDCE and an explicit one-off preference order in the ranking task. As such, there were only three ranks to be studied and there was a fixed order in best-best and best-worst but not in ranking. Due to only having three alternatives in the choice set, this study design allowed us to impose the same econometric model specification on a full preference order for all elicitation methods. As such, observed differences between elicitation methods can be directly attributed to the way in which the preferences were elicited. If wanting to use more than three alternatives, the econometric model specification no longer matches the data generation process of all elicitation methods and such a comparison is no longer possible. Alternative model specifications that match the data generation process such as the sequential best worst DCE model proposed in Lancsar et al. (2013) can then be used. Secondly, the findings from this sample may not generalize to other groups or respondents that complete DCEs. We collected data in a sample of taxpayers from an online panel in which the level of engagement in the choice tasks may differ from a sample of patients or clinicians. Lastly, although Hawkins et al. (2014) showed that the mere existence of a second choice did not change preferences of the first, it should be noted that we did not elicit single-best only. The direct comparison between single-best DCE and BBDCE would be a useful contribution to the health economics literature.

Other avenues for future research would be to further study the external validity of these elicitation methods, and to compare if any of these elicitation methods is more suitable for a particular type of research question or context. In some contexts, choices are framed in a positive manner and focus on the promotion of the best alternative, for example, preference for a treatment. In other contexts, choices are framed negatively and focus on prevention of the worst alternative, for example, when choosing to take medication to prevent a particular health state. Following this reasoning, it would also be interesting to elicit worst and second-worst preferences. Furthermore, it would be interesting to study the implications of our results for dual-response choice designs that are also argued to increase efficiency of data collection (Brazell et al., 2006).

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENT

CONFLICT OF INTEREST

None of the authors report conflicting interests.

DATA AVAILABILITY STATEMENT

Data and code are available upon request.

ETHICS STATEMENT

Monash University Human Research Ethics Committee approved the study.

ORCID

Samare P. I. Huls https://orcid.org/0000-0001-9864-037X
Bas Donkers https://orcid.org/0000-0002-0412-4276
Jemimah Ride https://orcid.org/0000-0002-1820-5499

ENDNOTES

[1] Best-worst scaling was first introduced by Finn and Louviere (Finn & Louviere, 1992). There are three types of best-worst scaling that differ in the respect that respondents either evaluate objects (case 1), attributes (case 2) or alternatives (also called multi-profile or BWDCE, case 3) (Flynn, 2010; Lancsar et al., 2013).

[2] As outlined in Bradley and Daly (1994), one can also set the scale parameter to be related to the random component of utility: $U_{ijt} = X'_{ijt}\beta_i + \varepsilon_{ijt} \cdot \lambda_s$. The position of the scale parameter does not change the results, only the interpretation.

[3] This linear specification of scale dynamics is chosen to reduce the number of parameters as no clear pattern could be observed from the data using a fully flexible specification. A quadratic specification led to a slight improvement of log-likelihood of a model with all betas and scale pooled but given the size of the improvement a linear specification was preferred for ease of interpretation.

[4] Results not reported, available on request.

REFERENCES

Akaichi, F., Nayga, R. M., & Gil, J. M. (2013). Are results from non-hypothetical choice-based conjoint analyses and non-hypothetical recoded-ranking conjoint analyses similar? *American Journal of Agricultural Economics*, *95*(4), 949–963. https://doi.org/10.1093/ajae/aat013

Beggs, S., Cardell, S., & Hausman, J. A. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics*, *17*(1), 1–19. https://doi.org/10.1016/0304-4076(81)90056-7

Ben-Akiva, M., Morikawa, T., & Shiroishi, F. (1991). Analysis of the reliability of preference ranking data. *Journal of Business Research*, *23*(3), 253–268. https://doi.org/10.1016/0148-2963(91)90033-T

Bradley, M., & Daly, A. (1994). Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation*, *21*(2), 167–184. https://doi.org/10.1007/BF01098791

Brazell, J. D., Diener, C. G., Karniouchina, E., Moore, W. L., Séverin, V., & Uldry, P.-F. (2006). The no-choice option and dual response choice designs. *Marketing Letters*, *17*(4), 255–268. https://doi.org/10.1007/s11002-006-7943-8

Brockner, J., & Higgins, E. T. (2001). Regulatory focus theory: Implications for the study of emotions at work. *Organizational Behavior and Human Decision Processes*, 86(1), 35–66. https://doi.org/10.1006/obhd.2001.2972

Caparrós, A., Oviedo, J. L., & Campos, P. (2008). Would you choose your preferred option? Comparing choice and recorded ranking experiments. *American Journal of Agricultural Economics*, 90(3), 843–855. https://doi.org/10.1111/j.1467-8276.2008.01137.x

Chang, J. B., Lusk, J. L., & Norwood, F. B. (2009). How closely do hypothetical surveys and laboratory experiments predict field behavior? *American Journal of Agricultural Economics*, 91(2), 518–534. https://doi.org/10.1111/j.1467-8276.2008.01242.x

Chapman, R., & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*, 19(3), 288. https://doi.org/10.2307/3151563

Cornelsen, L., Quaife, M., Lagarde, M., & Smith, R. D. (2020). Framing and signalling effects of taxes on sugary drinks: A discrete choice experiment among households in Great Britain. *Health Economics*, 29(10), 1132–1147. https://doi.org/10.1002/hec.4123

de Bekker-Grob, E. W., Donkers, B., Bliemer, M. C. J., Veldwijk, J., & Swait, J. D. (2020). Can healthcare choice be predicted using stated preference data? *Social Science & Medicine*, 246, 112736. https://doi.org/10.1016/j.socscimed.2019.112736

de Bekker-Grob, E. W., Ryan, M., & Gerard, K. (2012). Discrete choice experiments in health economics: A review of the literature. *Health Economics*, 21(2), 145–172. https://doi.org/10.1002/hec.1697

de Bekker-Grob, E. W., Swait, J. D., Kassahun, H. T., Bliemer, M. C. J., Jonker, M. F., Veldwijk, J., Cong, K., Rose, J. M., & Donkers, B. (2019). Are healthcare choices predictable? The impact of discrete choice experiment designs and models. *Value in Health*, 22(9), 1050–1062. https://doi.org/10.1016/j.jval.2019.04.1924

Delle Site, P., Kilani, K., Gatta, V., Marcucci, E., & de Palma, A. (2019). Estimation of consistent logit and probit models using best, worst and best–worst choices. *Transportation Research Part B: Methodological*, 128, 87–106. https://doi.org/10.1016/j.trb.2019.07.014

Dhar, R., & Wertenbroch, K. (2000). Consumer choice between hedonic and utilitarian goods. *Journal of Marketing Research*, 37(1), 60–71. https://doi.org/10.1509/jmkr.37.1.60.18718

Dyachenko, T., Reczek, R. W., & Allenby, G. M. (2014). Models of sequential evaluation in best-worst choice tasks. *Marketing Science*, 33(6), 828–848. https://doi.org/10.1287/mksc.2014.0870

Finn, A., & Louviere, J. J. (1992). Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy and Marketing*, 11(2), 12–25. https://doi.org/10.1177/074391569201100202

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), 259–267. https://doi.org/10.1586/erp.10.29

Fok, D., Paap, R., & Dijk, B. V. (2012). A rank-ordered logit model with unobserved heterogeneity in ranking capabilities. *Journal of Applied Econometrics*, 27(5), 831–846. https://doi.org/10.1002/jae.1223

Genie, M. G., Krucien, N., & Ryan, M. (2021). Weighting or aggregating? Investigating information processing in multi-attribute choices. *Health Economics*, 30(6), 1291–1305. https://doi.org/10.1002/hec.4245

Ghijben, P., Lancsar, E., & Zavarsek, S. (2014). Preferences for oral anticoagulants in atrial fibrillation: A best–best discrete choice experiment. *Pharmaco Economics*, 32(11), 1115–1127. https://doi.org/10.1007/s40273-014-0188-0

Giergiczny, M., Dekker, T., Hess, S., & Chintakayala, P. K. (2017). Testing the stability of utility parameters in repeated best, repeated best-worst and one-off best-worst studies. *European Journal of Transport and Infrastructure Research*, 17(4). https://doi.org/10.18757/ejtir.2017.17.4.3209

Gonzalez Sepulveda, J. M., Johnson, F. R., & Marshall, D. A. (2021). Incomplete information and irrelevant attributes in stated-preference values for health interventions. *Health Economics*, 30(11), 2637–2648. https://doi.org/10.1002/hec.4406

Hausman, J., & Ruud, P. A. (1987). Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics*, 34(1–2), 83–104. https://doi.org/10.1016/0304-4076(87)90068-6

Hawkins, G. E., Islam, T., & Marley, A. A. J. (2019). Like it or not, you are using one value representation. *Decision*, 6(3), 237–260. https://doi.org/10.1037/dec0000100. https://oce-ovid-com.eur.idm.oclc.org/article/01762424-201907000-00004/HTML

Hawkins, G. E., Marley, A. A. J., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). The best of times and the worst of times are interchangeable. *Decision*, 1(3), 192–214. https://doi.org/10.1037/dec0000012

Hess, S., & Rose, J. M. (2009). Allowing for intra-respondent variations in coefficients estimated on repeated choice data. *Transportation Research Part B: Methodological*, 43(6), 708–719. https://doi.org/10.1016/j.trb.2009.01.007

Higgins, E. T., Shah, J., & Friedman, R. (1997). Emotional responses to goal attainment: Strength of regulatory focus as moderator. *Journal of Personality and Social Psychology*, 72(3), 515–525. https://doi.org/10.1037/0022-3514.72.3.515

Krinsky, I., & Robb, A. L. (1986). On approximating the statistical properties of elasticities. *The Review of Economics and Statistics*, 68(4), 715–719. https://doi.org/10.2307/1924536

Krucien, N., Sicsic, J., & Ryan, M. (2019). For better or worse? Investigating the validity of best–worst discrete choice experiments in health. *Health Economics*, 28(4), 572–586. https://doi.org/10.1002/hec.3869

Lancsar, E., Louviere, J., Donaldson, C., Currie, G., & Burgess, L. (2013). Best worst discrete choice experiments in health: Methods and an application. *Social Science & Medicine*, 76(1), 74–82. https://doi.org/10.1016/j.socscimed.2012.10.007

Lancsar, E., Ride, J., Black, N., Burgess, L., & Peeters, A. (2021). Social acceptability of standard and behavioural economic inspired policies designed to reduce and prevent obesity. *Health Economics*, 31(1), 197–214. https://doi.org/10.1002/hec.4451

Lancsar, E., & Louviere, J. (2008). Conducting discrete choice experiments to inform healthcare decision making: A user's guide. *PharmacoEconomics*, 26(8), 661–677. https://doi.org/10.2165/00019053-200826080-00004

Laran, J., & Wilcox, K. (2011). Choice, rejection, and elaboration on preference-inconsistent alternatives. *Journal of Consumer Research*, 38(2), 229–241. https://doi.org/10.1086/659040

Louviere, J. J., Islam, T., Wasi, N., Street, D., & Burgess, L. (2008). Designing discrete choice experiments: Do optimal designs come at a price? *Journal of Consumer Research*, *35*(2), 360–375. https://doi.org/10.1086/586913

Louviere, J. J., Street, D., Burgess, L., Wasi, N., Islam, T., & Marley, A. A. J. (2008). Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. *Journal of Choice Modelling*, *1*(1), 128–164. https://doi.org/10.1016/S1755-5345(13)70025-3

Mansfield, C., Sutphin, J., & Boeri, M. (2020). Assessing the impact of excluded attributes on choice in a discrete choice experiment using a follow-up question. *Health Economics*, *29*(10), 1307–1315. https://doi.org/10.1002/hec.4124

Marley, A. A. J., & Louviere, J. J. (2005). Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology*, *49*(6), 464–480. https://doi.org/10.1016/j.jmp.2005.05.003

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, *4*(1), 103–120. https://doi.org/10.1080/0022250X.1975.9989847

Meloy, M. G., & Russo, J. E. (2004). Binary choice under instructions to select versus reject. *Organizational Behavior and Human Decision Processes*, *93*(2), 114–128. https://doi.org/10.1016/j.obhdp.2003.12.002

Palma, M. A. (2017). Improving the prediction of ranking data. *Empirical Economics*, *53*(4), 1681–1710. https://doi.org/10.1007/s00181-016-1169-2

Park, T., Loomis, J. B., & Creel, M. (1991). Confidence intervals for evaluating benefits estimates from dichotomous choice contingent valuation studies. *Land Economics*, *67*(1), 64–73. https://doi.org/10.2307/3146486

Petrolia, D. R., Interis, M. G., & Hwang, J. (2018). Single-choice, repeated-choice, and best-worst scaling elicitation formats: Do results differ and by how much? *Environmental and Resource Economics*, *69*(2), 365–393. https://doi.org/10.1007/s10640-016-0083-6

Punj, G. N., & Staelin, R. (1978). The choice process for graduate business schools. *Journal of Marketing Research*, *15*(4), 588–598. https://doi.org/10.1177/002224377801500408

Rose, J. M., & Bliemer, M. C. J. (2013). Sample size requirements for stated choice experiments. *Transportation*, *40*(5), 1021–1041. https://doi.org/10.1007/s11116-013-9451-z

Scarpa, R., Notaro, S., Louviere, J., & Raffaelli, R. (2011). Exploring scale effects of best/worst rank ordered choice data to estimate benefits of tourism in alpine grazing commons. *American Journal of Agricultural Economics*, *93*(3), 813–828. https://doi.org/10.1093/ajae/aaq174

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, *21*(4), 546–556. https://doi.org/10.3758/bf03197186

Soekhai, V., de Bekker-Grob, E. W., Ellis, A. R., & Vass, C. M. (2019). Discrete choice experiments in health economics: Past, present and future. *PharmacoEconomics*, *37*(2), 201–226. https://doi.org/10.1007/s40273-018-0734-2

Swait, J., & Louviere, J. (1993). The role of the scale parameter in the estimation and comparison of multinational logit models. *JMR, Journal of Marketing Research; Chicago*, *30*(3), 305. https://doi.org/10.2307/3172883

Train, K. (2003). *Discrete choice methods with simulation [online economics textbooks]*. SUNY-Oswego, Department of Economics. https://econpapers.repec.org/bookchap/oettbooks/emetr2.htm

Vanniyasingam, T., Cunningham, C. E., Foster, G., & Thabane, L. (2016). Simulation study to determine the impact of different design features on design efficiency in discrete choice experiments. *BMJ Open*, *6*(7), e011985. https://doi.org/10.1136/bmjopen-2016-011985

Vermeulen, B., Goos, P., & Vandebroek, M. (2011). Rank-order choice-based conjoint experiments: Efficiency and design. *Journal of Statistical Planning and Inference*, *141*(8), 2519–2531. https://doi.org/10.1016/j.jspi.2011.01.019

Xie, F., Pullenayegum, E., Gaebel, K., Oppe, M., & Krabbe, P. F. M. (2014). Eliciting preferences to the EQ-5D-5L health states: Discrete choice experiment or multiprofile case of best–worst scaling? *The European Journal of Health Economics*, *15*(3), 281–288. https://doi.org/10.1007/s10198-013-0474-3

Yangui, A., Akaichi, F., Costa-Font, M., & Gil, J. M. (2019). Comparing results of ranking conjoint analyses, best-worst scaling and discrete choice experiments in a nonhypothetical context. *The Australian Journal of Agricultural and Resource Economics*, *63*(2), 221–246. https://doi.org/10.1111/1467-8489.12292

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

APPENDIX A

TABLE A1 Literature overview

Authors (year)	Method				Criteria						Other	Context	Different preferences
	DCE	BB	BW	Rank	Coef./MRS	Scale	Scale dyn.	Efficiency	Self-rep.	Ext. validity			
Single method—within method comparison													
Ben-Akiva et al. (1991)				√	√	√						Transport	Yes
Delle Site et al. (2019)			√		√		√			√	Probit models	Transport	Yes
Hausman and Ruud (1987)				√	√	√						Transport	Yes
Hawkins et al. (2019)			√		√	√					Out-of-sample prediction	Marketing	No
Scarpa et al. (2011)			√			√	√					Environment	No
Lancsar et al. (2013)			√		√		√					Health	No
Multimethod—between method comparison													
Akaichi et al. (2013)	√		√		√					√	Proportion chosen, in-sample prediction	Agriculture	Yes
Caparrós et al. (2008)	√		√		√	√			√			Environment	No
Chang et al. (2009)	√		√		√					√	Market shares	Marketing	Yes
Petrolia et al. (2018)	√			√	√	√	√				Status-quo, attribute non-attendance	Environment	No
Xie et al. (2014)	√			√	√		√		√		Intraclass corr. Coeff.	Health	No
Yangui et al. (2019)	√		√	√	√					√	Intra-rater reliability	Agriculture	No
Multimethod—within and between method comparison													
Giergiczny et al. (2017)	√	√	√		√	√			√			Transport	Yes
Hawkins et al. (2014)	√		√		√						Proportion chosen, in-sample prediction	Marketing	No
Krucien et al. (2019)	√		√		√	√						Health	Yes
This study		√	√	√	√	√	√	√		√		Health	Yes

Abbreviations: BB, best-best; BW, best-worst; DCE, discrete choice experiment.

TABLE A2 Questions respondent-reported criteria, per criterion and per method

Question [response options]	BW	BB	Rank
Difficulty [easy, neither, difficult]			
Thinking about the choices you just made, how did you find the task of choosing the best option in each scenario?	√	√	
Thinking about the choices you just made, how did you find the task of choosing the worst option in each scenario?	√		
Thinking about the choices you just made, how did you find the task of choosing the second best option in each scenario?		√	
Thinking about the choices you just made, how did you find the task of ranking the 3 options in each scenario?			√
Preference [yes, no]			
Instead of choosing best and then worst in each scenario, would you have preferred to choose the first best (like you did) and then instead of choosing worst of the remaining two options, choose best of the remaining two options?	√		
Instead of choosing best and then worst in each scenario, would you have preferred to rank the options from best to worst?	√		
Instead of choosing best and then second best in each scenario, would you have preferred to choose the first best (like you did) and then instead of choosing second best of the remaining two options, choose worst of the remaining two options?		√	
Instead of choosing best and then second best in each scenario, would you have preferred to rank the options from best to worst?		√	
Instead of ranking the options in each scenario from the best to worst, would you have preferred to choose the first best and then choose second best of the remaining two options?			√
Instead of ranking the options in each scenario from the best to worst, would you have preferred to choose the first best and then choose worst of the remaining two options?			√

Abbreviations: BB, best-best; BW, best-worst.

TABLE A3 Characteristics of participants per arm

Variables	Best-worst $N = 997$	Best-best $N = 926$	Ranking $N = 1012$
Age, years, mean (SD)	44.0 (14.3)	43.6 (14.4)	44.1 (14.1)
Gender, female (%)	447 (44.8)	408 (44.1)	449 (44.4)
Employment status, employed (%)	843 (84.6)	798 (86.2)	855 (84.5)
Education, n (%)			
Never completed (high) school	88 (8.8)	89 (9.6)	92 (9.1)
High school	133 (13.3)	125 (13.5)	135 (13.3)
Technical and Further Education (TAFE)	237 (23.8)	221 (23.9)	210 (20.8)
Associate diploma	83 (8.3)	76 (8.2)	81 (8.0)
University	443 (44.4)	414 (44.7)	493 (48.7)
Other	13 (1.3)	1 (0.1)	1 (0.1)
Self-assessed health, n (%)			
Excellent	106 (10.6)	96 (10.4)	102 (10.1)
Very good	340 (34.1)	334 (36.4)	354 (35.0)
Good	376 (37.7)	341 (36.8)	378 (37.4)
Fair	126 (12.6)	127 (13.7)	147 (14.5)
Poor	49 (4.9)	28 (3.0)	31 (3.1)
Duration in minutes, mean (SD)	40,7 (56.1)	37,4 (44.9)	36,9 (33.4)
Duration in minutes, median	28.4	27.5	29.1