A systematic review comparing the acceptability, validity and concordance of Discrete Choice Experiments and Best Worst Scaling for eliciting preferences in healthcare

Jennifer A. Whitty[1]

A. Sofia Oliveira Gonçalves[1]

Affiliations

1. Health Economics Group, Norwich Medical School, University of East Anglia, Norwich, Norfolk, UK

Conflicts of Interest

JW has previously published both DCE and BWS studies in this field. The authors are not aware of any potential conflicts of interest related to the review

Sources of Funding

No funding support was received for this study.

Corresponding author

Name: Jennifer Whitty

Address: Health Economics Group, Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich UK NR4 7JT

Email: Jennifer.whitty@uea.ac.uk

Phone: +44 (0)1603 593536

Acknowledgements

Sofia Goncalves performed the searches, screened and extracted papers, assessed study quality, drafted the review, and approved the final version of the review prior to submission.

Jennifer Whitty coordinated and edited the review, reviewed papers for inclusion, made an intellectual contribution, approved the final version of the review prior to submission and is the guarantor of the review.

Word count: 8,281 (+ 1 figure, 2 tables, plus supplementary material)

A systematic review comparing the acceptability, validity and concordance of Discrete Choice Experiments and Best Worst Scaling for eliciting preferences in healthcare

ABSTRACT

Objective: To compare the acceptability, validity and concordance of discrete choice experiment (DCE) and best-worst scaling (BWS) stated preference approaches in health.

Methods: A systematic search of EMBASE, Medline, AMED, PubMed, CINAHL, Cochrane Library and EconLit databases was undertaken in October to December 2016 without date restriction. Studies were included if they were published in English, presented empirical data related to the administration or findings of traditional format DCE and object, profile or multiprofile-case BWS, and were related to health. Study quality was assessed using the PREFS checklist.

Results: Fourteen articles describing twelve studies were included, comparing DCE with profile-case BWS (9 studies), DCE and multiprofile-case BWS (1 study), and profile and multiprofile case BWS (2 studies). Although limited and inconsistent, the balance of evidence suggests that preferences derived from DCE and profile-case BWS may not be concordant, regardless of the decision context. Preferences estimated from DCE and multiprofile-case BWS may be concordant (single study). Profile- and multiprofile-case BWS appear more statistically efficient than DCE, but no evidence is available to suggest they have a greater response efficiency. Little evidence suggests superior validity for one format over another. Participant acceptability may favour DCE which had a lower self-reported task difficulty and was preferred over profile-case BWS in a priority-setting but not necessarily other decision contexts.

Conclusion: DCE and profile-case BWS may be of equal validity but give different preference estimates regardless of the health context; thus, they may be measuring different constructs. Therefore, choice between methods is likely to based on normative considerations related to coherence with theoretical frameworks and on pragmatic considerations related to ease of data collection.

Keywords: *best worst scaling, maximum difference, discrete choice experiment, conjoint analysis, stated preference, pairwise choices, preference measurement, validity*

KEY POINTS

- It is imperative to consider patient and public preferences in healthcare decision-making. Understanding the merits of different preference elicitation formats is important to support their accurate measurement.
- This evidence-based systematic review suggests that Discrete Choice Experiments (DCE) and profile-case Best-Worst Scaling (BWS) formats may be of equal validity but give different preference estimates in the health setting, suggesting these methods are measuring different constructs and are therefore not interchangeable.
- This appears to the case regardless of whether the choice context relates to personal preferences for treatment, or social preferences for prioritising the treatment of others.
- However, there may be some pragmatic advantages to one format over another. For example, BWS may generally require a smaller participant number to estimate precise preferences suggesting greater statistical efficiency, and DCE may be more acceptable to participants suggesting greater response efficiency in a priority-setting context.

1 INTRODUCTION

Understanding the preferences of patients and the public around the processes and outcomes of healthcare is paramount to designing and evaluating interventions and services. In the healthcare setting, the concept of using a stated preference method to value a good or service and inform its design through understanding its preferred characteristics has been extended to also include indicating the relative value of different outcomes from healthcare. Applications of stated preference methods in health have included the valuation of different health states to derive utility weights (health state valuation, HSV), considering the trade-offs people are willing to make between treatment risks and benefits to estimate maximum acceptable risk (MAR) thresholds, and valuing process outcomes [1, 2]. Further, use of stated preference approaches to estimate value judgements in priority setting and potentially incorporate these in decision-making (e.g. through multi-criteria decision analysis) is also growing [3-6]. Given the popularity of stated preference methods in the healthcare setting, research to increase our understanding of the relative merits and limitations of different stated preference elicitation formats is needed to inform their methodological development.

Approaches utilised to elicit and quantify preferences in a healthcare setting include the Time Trade-Off (TTO) and Standard Gamble (SG) techniques which have routinely been applied in health state valuation [7], the Person Trade-Off (PTO) which has been applied to measure social value judgements (e.g. [8]), and Contingent Valuation (CV) which has been used to measure willingness to pay for a health service (e.g. [9]). Although each of these approaches have merit [10], they are limited in that they are only able to measure preferences according to the trade-offs inferred between two characteristics. This is typically health-related quality of life (HRQoL) and time lived in a health state (TTO) or risk of death (SG), a health gain or treatment characteristic (e.g. age) and the number of people treated (PTO), or a health service profile and out of pocket cost (CV). There has been growing interest in the application of alternative preference elicitation approaches which are capable of eliciting trade-offs between more than two characteristics; in particular, the Discrete Choice Experiment (DCE) and Best Worst Scaling (BWS) [2, 11, 12].

The DCE, which is also referred to as choice-based conjoint analysis, has become popular as a to elicit preferences for alternative interventions, services or outcomes in the health care sector [11, 12].

DCEs are an ordinal stated preference method based on random utility theory developed by Thurstone [13] and operationalised by McFadden [14]. As such, they assume that each individual attaches a latent or unobserved "utility" to each choice alternative [15], which is a function of the utility associated with unique "attributes" of that alternative [16]. In the traditional DCE format, participants are asked to choose their preferred option from two or more alternative profiles described by a number of attributes, where the levels of each attribute vary across alternatives [17].

More recently, another ordinal method has been applied to elicit preferences for health and healthcare: best-worst scaling (BWS). Developed by Louviere and Woodworth (1990) and Finn and Louviere (1992) [18, 19], this elicitation method refers to choice tasks which contain at least three alternatives and in which participants need to choose the "best" and the "worst" alternative [20]. Best could also be considered as most preferred or most important and worst as least preferred or least important. BWS tasks can take the form of three different cases: "object" (case 1), "profile" (case 2) and "multiprofile" (case 3) [21]. A full description and example for each of these three cases is provided by Flynn 2010 [22] and a review of their application in healthcare has been provided by Cheung et al. [23]. Briefly, in the BWS object case, participants are asked to choose the best and worst option among different whole objects, which are not separated into attributes (for an application in health, see for example [24]). In the profile case, the choice set has the structure of a single profile and shows the level of each attribute. The participant does not need to take the overall profile's value into account, but needs to consider the attribute-levels which define it, and select the best and the worst one. Thus, as the attribute levels are part of a complete profile, there is no explicit choice trade-off *per se*. This has been the most commonly applied approach in health [23]; see for example [25-30]. The multiprofile-case asks participants to choose the best and the worst profile in a choice set; hence, the choice trade-off is between different alternative profiles consisting of a number of attribute levels (see for example [31, 32]). Thus, the multiprofile-case BWS is similar to a traditional format DCE [2], but must present at least three profiles in each choice set and elicits more information than a DCE, by asking for the 'worst' as well as the 'best' profile. If 'best' and 'worst' are elicited repeatedly (best, worst, next best, next worst etc), a full preference ranking can be obtained. Whilst a similar approach to maximising data collection is also possible with DCEs (a researcher could ask best, next best etc), it has not generally been undertaken in the traditional DCE context.

The number of published studies which estimate preferences using BWS methods lags behind the ones that use DCEs which is consistent with the more recent introduction of BWS to the health context. Nevertheless, both DCE and BWS formats continue to gain popularity for estimating preferences in the health sector [3, 12, 23, 33][34]. Both methods have typically been operationalised in health using a similar random utility theoretical basis, enable the valuation of health care treatments and outcomes, and support the participation of patients or the public in the decision-making process and the consideration of their preferences alongside those of different stakeholders [34]. Nevertheless, despite the growth in studies using either approach, there is little guidance available on which approach is best to use and in what circumstances in the healthcare setting.

DCEs and BWS make different assumptions about the participants underlying psychological decision model and choice behaviour. However, they both intend to estimate preferences for health and healthcare for the same purpose. Researchers have argued that BWS approaches may be superior to traditional format DCEs on the basis that (i) each BWS choice derives more information about the participant's ranking of alternatives; (ii) BWS exploits an individuals' propensity to easily recognise and answer the two most extreme options in a choice set; and (iii) BWS tasks may be easier for the general population to understand [35][20, 22, 36]. Others have claimed that profile-case BWS can be easily understood in comparison with the DCE or multiprofile-case BWS, since each scenario only entails one hypothetical profile to be evaluated [37]. However, little empirical data has been provided to support these arguments. Conversely, arguments favouring more traditional DCE formats have highlighted that profile-case BWS, which has been the most commonly applied BWS case in health to date, requires participants to make choices within and not between profiles [38]. Thus, it has been argued that profile-case BWS does not conform to the welfarist requirement for a stated preference method to imply trade-offs, and that the profile-case BWS elicits 'values' and not 'preferences' [29, 38]. Nevertheless, health economic evaluation is currently dominated by an extra-welfarist paradigm; therefore, this might be argued to be a theoretical consideration but not a pragmatic limitation for the profile-case BWS format.

Empirical data evidencing the fundamental properties of measurement instruments, such as participant acceptability and validity, are crucial to inform their comparative merits. Although there are already several published reviews on DCE [1, 11, 12] and BWS [2, 39] in a health context, these reviews

have examined the two approaches separately. There has been no comprehensive review of studies which directly compare both preference elicitation methods to guide researchers on the relative merits of the two approaches. To address this gap, we undertook a systematic review of the published literature to compare the acceptability, validity**,** and concordance of DCE and BWS elicitation formats when applied to elicit preferences or values in a health setting.

2 METHODS

The recommendations of the *Preferred Reporting Items for Systematic Reviews and Meta-analysis – PRISMA*) were followed as a framework for the systematic review. We acknowledge the debate on terminology and whether DCE and BWS formats elicit "preferences" or "values" [29, 38]. However, for simplicity, we will refer to the output of the DCE and BWS formats as "preferences" in this review.

2.1 Eligibility Criteria

This systematic review included studies published in English, which compared traditional format DCEs with BWS (object, profile or multiprofile case) or compared different cases of BWS approaches in the health setting. For the purpose of this review, traditional format DCEs were defined to include DCEs presenting two or more profiles, DCEs presenting a single profile against a constant comparator, and DCEs presenting a single profile and seeking a dichotomous (e.g. accept/reject) choice. Studies presenting qualitative and/or quantitative empirical data related to the administration or findings of DCE and BWS or of two or more different cases of BWS were included. Non-health related studies, studies that did not present original empirical data, and studies published as an abstract only were excluded.

2.2 Literature Search

A search of the databases *EMBASE*, *Medline*, *Allied and Complementary Medicine Database* (*AMED)*, *PubMed*, *Cumulative Index to Nursing and Allied Health Literature* (*CINAHL)*, *Cochrane Library* and *EconLit* was conducted between October and December 2016 without date restriction. The search strategy was based on the approach used by Ryan & Gerard [11] to identify DCE studies, using

the following terms: ‘discrete choice experiment’, ‘DCE’, ‘discrete choice model’, ‘discrete choice modelling’, ‘conjoint analysis’, ‘conjoint study’, ‘conjoint choice experiment’, ‘stated preference’, ‘pairwise choices’ and ‘paired comparisons’. For BWS the terms included: ‘best worst scaling, ‘best-worst scaling’, ‘BWS’, 'maximum difference', ‘maxdiff’ ‘best worst survey’ and ‘best worst scoring’. The search strategies are available in the online supplementary appendix. In addition, manual searches of the reference lists of included publications and of online records were conducted.

2.3 Screening for eligibility, data extraction and quality review

Titles and abstracts of all citations were reviewed for eligibility. Duplicates were removed. Potentially eligible papers were retrieved and reviewed for inclusion by two researchers, with any disagreements resolved by consensus.

Data were extracted with special consideration to the comparative acceptability and validity of the methods and the concordance of their findings. The comparative acceptability and validity of the methods indicate the quality of the data collected. Acceptability is the extent to which a choice format is considered to be acceptable to participants. It is therefore likely to impact other important considerations for a stated preference method including feasibility of administration and response efficiency. Response efficiency is the ease with which high quality data giving precise estimates can be collected using the method, and has been identified as being an important and often overlooked aspect of stated preference study design [40]. This is in contrast to statistical efficiency which relates to the efficiency with which the underlying statistical design gives precise preference estimates. To evaluate acceptability for each of the DCE or BWS formats, we extracted data (where available) related to the response rate, completion time for the task(s), self-report task difficulty, and participant preferred method. Data relating to acceptability was considered to be more easily extractable than data relating to feasibility or response efficiency as this is more likely to be reported upon in published studies.

Assessing validity is of great importance, since interpretation of the findings of both DCE and BWS stated preference methods rely on answers to hypothetical choices [1]. Internal validity is particularly relevant as if it was violated, external validity would also not hold [41]. To evaluate the internal validity of the methods, data indicating compliance with the following axioms of utility maximisation were collected from the studies: (i) continuity; (ii) transitivity; (iii) monotonicity; (iv)

consistency (stability); and (v) completeness [38, 42]. Data related to convergent validity, comparing the DCE/BWS method with any other relevant internal or external measure, were also extracted.

Concordance indicates whether the two methods are measuring the same construct and lead to a decision model that would predict the same choice behaviour. To evaluate concordance, we extracted data on both the similarity of preferences and on decision certainty. Data were sought indicating the similarity of preferences estimated by the two methods, as assessed by the direction of preference, rank order of preference, correlation between preference estimates from the methods, or the absolute size of preferences after rescaling of preference weights (for example, via estimates of marginal rates of substitution). If preferences are similar, it suggests the methods are measuring the same construct and can be used interchangeably to predict decision-making. Conversely, if preferences are different, it suggests the methods are either measuring different constructs or are measuring the same construct but with different levels of validity. Data were also sought on the comparative certainty of choices made by participants for the two different choice tasks, to provide insight into decision (un)certainty and whether the methods are likely to lead to statistically efficient (precise) preference estimates for a given sample size. This was assessed through exploration of the comparative scale of the error term for each logit choice model, which is inversely related to the error variance of the model. A high scale suggests low error variance and low randomness in decision-making (i.e. high decision certainty), whilst low scale suggests the converse, If the scale of the error term is similar between models, this suggests decision certainty is similar, and similar efficiency in data collection. Thus, similar sample sizes would be needed for a given level of precision in the preference estimates.

Overall study quality was assessed based on the PREFS quality assessment checklist [43], which is suitable for application across both DCE and BWS methods. This checklist comprises five domains (Purpose, Responders, Explanation, Findings and Significance) and each one can only take the value of 1 (acceptable) or 0 (not acceptable). Review findings are reported descriptively; no attempt at meta-analysis was made given the heterogeneity of the studies.

3 RESULTS

3.1 Article Inclusion

A total of 559 articles were identified through the searches (Figure 1). A further three article were identified through manual searches. After removing duplicates, 248 articles were reviewed for title and abstract. A total of 219 did not meet the inclusion criteria and were excluded from this review. The remaining 29 articles were screened in full, which resulted in the exclusion of 15 more articles. Fourteen articles met the inclusion criteria. The 14 articles reported 12 separate studies, with two articles by Whitty et al. reporting a think aloud pilot and main data collection and analysis for the same study [17, 44] and two articles by Netten, Potoglou et al. reporting data based on the same social care study [45, 46].

3.2 Overview of studies

Table 1 describes the main characteristics of the 12 studies included in this review. All studies were published since 2010, an observation that is consistent with the recent trend of using DCE and BWS methods for evaluating health care preferences [34]. Furthermore, all were conducted in high income countries (according to the World Bank definition) and all except two were conducted only in English speaking countries. Ten studies compared DCE with BWS [17, 38, 44-53]:, nine with BWS profile-case and one with BWS multiprofile-case [53]. Two studies compared BWS profile to BWS multiprofile-case [37, 54].

Five studies were undertaken to value outcomes: four studies in the context of health state valuation and one in social care outcome valuation (Table 1). These valuation studies compared DCE with profile (four studies) or multiprofile (one study) case BWS. Four of the valuation studies used general population samples who did not necessarily have experience of the health states being valued; the fifth valued a condition-specific instrument using the preferences of patients with that condition (glaucoma).

Four studies elicited preferences for treatment options; three compared DCE with BWS profile-case and one compared profile and multiprofile BWS. These studies recruited a wide mix of participants: patients with experience of the condition in question, caregivers, potential patients who had not necessarily experienced the condition but were in a higher risk group, and general population. Two

studies elicited priority-setting preferences of the general population or experts, both compared DCE with BWS profile case. One study elicited nurse job preferences.

All studies presented comparative data based on the main survey findings from their preference study. In addition, four studies presented pilot or qualitative findings related to the comparison of two methods from multiple pilot studies [47], pretest interviews [45, 50], or a pre-test Think Aloud study [44].

3.3 Study design and methods

3.3.1 Recruitment and data collection

Each study used the same data collection methods for both the DCE and BWS or across different BWS cases. Recruitment methods of interviewees were diverse, and included "House-to-house recruitment", "Internet panel" and "Advertisement". Most surveys were conducted with the support of a computer, either using computer-aided personal interviews or online surveys. Importantly, given the intent of comparing DCE and BWS data, ten of the twelve studies clearly indicated they used the same participant cohort to complete both DCE/BWS tasks (Table 1). Only four of these randomised which task (DCE or BWS) was seen first.

3.3.2 Choice Sets' Design and Estimation Procedure

The studies comparing a DCE with a BWS task most commonly used DCE choice sets presented generically in pairs (5 studies; Table 2); the others used a constant comparator or a single profile with a categorical choice. Studies considering multiprofile BWS all used triplet BWS choice sets. The only study comparing DCE and multiprofile-case BWS used pairs not triplets for DCE, even though the BWS task used triplets [53].

The studies included between 5 and 12 attributes in their DCE/BWS design. All studies used the same attributes for the DCE and BWS tasks, with the exception of Netten, Potoglou et al. [45, 46], who used 9 social care outcome attributes for their BWS and split these across two separate DCE tasks (with duplication) to reduce participant burden.

The combination of attributes with levels frequently leads to a large amount of possible scenarios (full factorial design). Fractional factorial designs avoid an excessive number of choice tasks

for participants. They require the selection of a number of choice sets which allows main effects and (optionally) potential interaction effects between attributes to be elicited [1]. All studies considered for this review used fractional factorial designs and some blocked the design across participants; although, van Dijk and colleagues may have covered the full factorial with their design (not explicitly stated) [52]. The majority of studies used an orthogonal main effects design for both DCE and BWS studies. Several studies used software packages to create a fractional factorial design: SAS [45, 46, 49, 53], NGENE [17, 44] and Sawtooth [52, 54]. Only seven of the twelve studies appeared to use the same underlying statistical design matrix to specify the choice profiles used for both DCE/BWS tasks (Table 2).

Most studies used the same statistical modelling frameworks for estimating preferences using choice data from the DCE and BWS tasks (Table 2). Although the studies included in this review are relatively recent (2010 onwards), they generally used logit and probit models for data analysis. Three studies used a Mixed Logit (MXL) and/or a Generalised Multinomial Logit (GMNL) model, which require less restrictive assumptions than conditional or multinomial logit and allow preferences to vary across participants. Three studies stated they accounted for interaction effects between attributes in their analysis [17, 45-47]; although, only for the DCE and not for the BWS data.

3.4 Acceptability

Data providing insights into the acceptability of the approaches for participants were provided in terms of the response rate, task completion time, task difficulty, and method preferred by participants.

3.4.1 Response rate

Few studies reported response rates. However, 10 of the 12 studies used the same participants to complete the DCE and BWS (or different cases of BWS) tasks (Table 1); thus, the response rates would not differ by method. The two studies that used different participants for each task did not report response rate [50, 51].

3.4.2 Completion time

Only three studies reported completion time by task. Two compared DCE with profile-case BWS. Severin et al. reported a similar completion time for the DCE and profile-case BWS surveys, which were completed by different participants [51]. On the other hand, van Dyke et al. reported a

shorter median completion time for the DCE than profile-case BWS tasks, which were completed by the same participants (with task order randomised) [52]. The third study reported a shorter mean response time to complete 12 DCE paired choice tasks (7.8 minutes) than for 12 multiprofile-case BWS triplet tasks (10.3 minutes) [53].

3.4.3 Task difficulty

Four studies reported the comparative difficulty of the DCE and profile-case BWS tasks. Findings were diverse. Potoglou et al. reported a similar number of participants were excluded from the DCE and profile-case BWS analysis as a result of combined self-report and interviewer reported difficulty with the survey [46], suggesting a similar level of difficulty completing both tasks. The self-reported median completion difficulty was also similar for DCE and profile-case BWS in van Dyke et al. [52]. However, self-report data from two studies suggest participants found the profile-case BWS task to be more difficult than the DCE task. In Severin et al., a greater proportion of participants found the profile-case BWS difficult to understand than for DCE; however, a similar proportion of participants reported both tasks to be difficult or very difficult to answer [51]. In Whitty et al. a greater proportion of participants found the profile-case BWS difficult to complete than for DCE [17]. This is consistent with the preceding think aloud study, in which some participants were observed to find it challenging to choose a least important attribute/level in the BWS task, when the least important might still be perceived as important, or as not important at all, for a funding decision [44]. In the only study comparing DCE and multiprofile-case BWS, more participants (22%) reported severe difficulty choosing from three (BWS) than two (DCE) health states (16%) [53]. No study reported the DCE to be more difficult to complete than the BWS task.

3.4.4 Participant preferred task format

Two studies asked participants which task they preferred; both compared DCE with profile-case BWS. Participants in the pre-test interviews carried out by Janssen et al. did not display a predilection for either DCE or profile-case BWS tasks: some participants found DCE better since it looked similar to everyday decision-making, while others preferred profile-case BWS because it contained less information [50]. Conversely, approximately three quarters of participants in Whitty et al. (75% in think aloud study, 72.6% in main preference elicitation survey) indicated they preferred the DCE over profile-

case BWS task [17, 44]. Participants in the think aloud study who preferred DCE stated this was because the DCE allowed a comparison of full profiles; whereas, those preferring profile-case BWS did not want to choose a single undesirable characteristic that was part of a whole package, or perceived BWS to be less ethically conflicting or burdensome [44].

3.5 Validity

Selected studies provided evidence indicating compliance with the axioms of utility maximisation supporting the comparative internal validity of the methods.

3.5.1 Axioms of utility maximisation

Continuity

Continuity of preferences assumes compensatory decision-making [38, 55]. Two studies provided comparative evidence of participants trading between attribute levels, indicating continuity of preferences. In their think aloud study, Whitty et al. reported evidence of trading to be strongly observed for DCE and weakly observed for profile-case BWS, with some participants showing a lack of variation in their best or worst choices [44]. Krucien et al. computed a lexicographic score for each participant to indicate their level of trading between attributes, varying from 0% (never selects a given attribute as best or worst) to 100% (always selects the alternative with the highest level of a given attribute (in DCE) or always selects a given attribute as best or worst) [38]. Dominant preferences were assumed for a given attribute when the lexicographic score exceeded 50%. More participants exhibited dominant preferences for a single attribute in the profile-case BWS task (23.5%) than DCE (16.6%; p=0.047), suggesting greater trading occurred for DCE than for BWS. In both tasks, two thirds of those categorised as having dominant preferences did so for the same attribute.

Monotonicity

As long as commodities are “goods” and not “bads”, participants are assumed to prefer more than less. Thus, monotonicity implies that utility functions are increasing with an increasing quantity of a desirable attribute or a decreasing quantity of an undesirable attribute [38, 56]. For DCEs monotonicity can be tested by observing whether a participant chooses a “dominant” profile, that is, one that is more desirable (dominates) the other on one or more attributes and is at least as good on all others [38].

However, for profile-case BWS, testing for monotonicity is more challenging than for DCE, since a priori expectations indicating the expected best (dominant) or worst (dominated) attribute level in a profile may not be apparent, and so tests have not been widely established.

Two studies comparing DCE and profile-case BWS conducted dominance tests. The most comprehensive assessment was made by Krucien et al., representing the only study to attempt to compare monotonicity across methods. They reported that 73% of participants fully satisfied and 2% fully failed a conventional monotonicity test for the DCE task (consisting of a dominance test for each of 5 tasks) [38]. This compared to 0% fully satisfied and 42% fully failed a modified monotonicity test for BWS. However, the BWS dominance test consisted of 6 tasks – making failure purely by chance more likely than for the DCE test. Moreover, the BWS test relied on an assumption of dominance based on the level descriptors alone regardless of the health state domain concerned. For example, 'some difficulty' for one domain was always assumed to be better than 'quite a lot of difficulty' for another domain, regardless of which domain was being considered. This seems a strong assumption, as it requires that the domains themselves are of equal value – which is unlikely to be the case. Indeed, an assumption that domains are *not* of equal value is a key driver for the derivation of health state valuations for multi-attribute instruments.

In the second study, Severin et al. reported that one participant failed to select a dominant alternative in a DCE choice set; however, there was no comparative assessment of monotonicity for the BWS task [51]. A third study did not conduct a test of monotonicity, but observed a lack of monotonicity for the estimated preference weights across the levels for two attributes in the DCE but not BWS analysis [49]. However, the DCE was a dichotomous yes/no choice format and thus did not require a trade-off to be made across profiles [49].

The only study comparing DCE and multiprofile-case BWS found a high and similar proportion of participants passed tests for monotonicity; 99% chose a dominant choice set in the DCE tasks, and 98% (97%) chose a dominant (dominated) health state as most (least) preferred in BWS tasks [53].

Consistency

Three studies comparing DCE and profile-case BWS assessed consistency (also referred to as stability), by asking participants to complete one or more repeat choice tasks [17, 38, 44, 52]. Preferences were deemed to be consistent if participants chose the same alternative in successive choice tasks. Such tests have been criticised since individuals might remember their answer, or preference reversal might occur due to very similar alternatives which could be considered as close substitutes. Nevertheless, tests of consistency provide an indicator of internal validity and data quality.

Van Dyke et al. reported a similar level of consistency between DCE and profile-case BWS (>96% of participants were consistent for both); although, consistency was not clearly defined for this study [52]. Although they reported lower overall levels of consistency than van Dyke et al., both Whitty et al. and Krucien et al. reported greater consistency for DCE than profile-case BWS [17, 38, 44]. However, the probability of being consistent to the repeat choice task(s) by chance alone is higher for the DCE task (50% chance in Whitty et al. and 25% chance in Krucien et al.) than for the BWS task (2.4% to 14.3% chance in Whitty et al. and 5% chance in Krucien et al.) [17, 38]. Whitty et al. reported lower levels of consistency for the worst than for the best choices in the profile-case BWS task [17].

The only study comparing DCE and multiprofile-case BWS found greater consistency for the DCE task than for the multiprofile BWS task. Xie et al. reported a higher intra-rater agreement for DCE (intraclass correlation coefficient ICC 0.53) across three repeat tasks than for BWS (ICC 0.45) across two repeat tasks [53]. They reported similar rates of inconsistency in BWS for best and worst choices.

Completeness

If participants are aware of their preferences and they are easily uncovered by the researcher they are deemed to be complete. Technically, the axiom of completeness represents the ability of participants to make a choice according to a rank ordering of available options [38]. Krucien et al. reported the DCE and profile-case BWS to perform similarly in an indirect test for completeness of preferences [38]. However, they also found ranking from the two methods to be uncorrelated (Kendall correlation -0.222, p=0.146), concluding that the DCE and profile-case BWS "perform equally well (in terms of completeness) but in different ways" [38]. No other study reported comparative data supporting completeness.

Transitivity

Transitivity considers that if choice A is preferred to choice B and choice B is preferred to choice C, then choice A should be preferred to choice C [55]. Although this axiom is a central test of rationality (alongside completeness) [56], it has not been widely applied in the DCE context [1]. No study in this review provided data supporting the axiom of transitivity.

3.5.2 Convergent validity (internal or external)

Only two studies reported data indicating convergent validity for the DCE/ BWS method against another internal measure. Xie et al. reported a high and similar level of agreement in the ranking of health states for both the DCE and multiprofile-case BWS when compared to visual analogue scale (VAS) valuations [53]. Weernink et al. reported aggregate level utility values derived from profile and multiprofile-case BWS to be highly correlated with those derived from TTO or VAS (Pearson correlation R^2 0.95 to 0.98, $p<0.001$) [54]. This was also the case at the individual (within person) analysis level (Pearson correlation R^2 0.56-0.68, $p<0.001$). However, better differentiation was observed for closely related treatment profiles for both BWS methods than for TTO or VAS at the individual level. The study did not discriminate between the BWS methods. In a third study, Netten et al. undertook a Time Trade Off (TTO) exercise in addition to a DCE and profile-case BWS [45]. However, their aim was to supplement the BWS (to derive utility weights) and not to compare the convergence of DCE or BWS with TTO. None of the studies reported data indicating the external validity of the methods.

3.6 Concordance of DCE and BWS findings

Preferences obtained through DCEs and BWS cannot be directly compared, since the underlying scales of the analytic model are different [51]. Studies used a range of approaches to adjust findings for scale, thus allowing a comparison of concordance across methods. Netten, Potoglou et al. and Severin et al. rescaled the results obtained from the DCE and BWS experiment choice models, examining the relative size of the differences [45, 46, 51]. Instead of rescaling the obtained results, Janssen et al.

compared values using the correlation measure Spearman's rho [50]; whilst, Whitty et al. compared values after rescaling and computed a Pearson correlation coefficient [17].

3.6.1 Comparison of DCE and profile-case BWS findings

Concordance of preference estimates

Most studies identified in this review compared DCE and profile-case BWS. Findings on concordance of preferences are mixed; however, overall they suggest a low level of concordance between the two methods, particularly if the preference weights are used to estimate marginal rates of substitution. Most studies reported comparable preference patterns in terms of direction of preference and (in most cases) rank order of preference for gains in different attributes for the DCE and profile-case BWS, with correlation coefficients (where reported) of approximately 0.9 or above [46, 47, 49-52]. However, several of these studies reported lower concordance of ranking across the mid-ranked as opposed to the high or low ranked attributes [49, 51]. Moreover, several of these studies also indicated that whilst the pattern of preferences were broadly similar, the estimated preferences differed in "detail" in terms of the comparative size of the preference weight estimates [46, 51], and in particular thresholds estimated (such as the Minimal Acceptable Risk, MAR) could differ [52].

Two studies reported markedly different findings between the DCE and profile-case BWS. In a priority-setting application, Whitty et al. reported poor correlation between DCE and BWS weights (Pearson correlation coefficient 0.286, p=0.282), and the relative size of the preference weights and rank ordering differed between methods implying their use might lead to different priorities assigned [17]. More recently, Krucien et al. also reported poor concordance in a health state valuation application between rescaled parameter estimates (Spearman correlation coefficient 0.272 to 0.610; $p<0.05$) and a systematic bias between DCE and profile-case BWS rescaled coefficients [38]. They also reported less discrimination between level weights within attributes for BWS than for DCE, implying a difference between methods if they were used to derive utility weights for use in economic evaluation.

Decision (un)certainty

Studies that explored scale differences between the DCE and profile-case BWS models generally reported higher scale and therefore lower error variance suggesting greater decision consistency and potentially lower decision uncertainty with profile-case BWS than DCE [47, 48, 52],

with the exception of Whitty et al. who found the converse [17]. The lower variance may also have been an artefact of the greater amount of choice observations collected for BWS (two rather than one for each choice task). Lower error variance in the model would lead to more precise preference estimates, and therefore a smaller sample might be required to complete the preference study. Krucien et al. found no difference in scaling between DCE and profile-case BWS [38]. Therefore, they attributed their observations of difference in preferences to differences in the accuracy rather than the precision of the methods. Comparisons of error variance and scale are dependent on the characteristics of participants in addition to any difference in the methods. Therefore, any valid comparison of decision certainty between methods should ideally be undertaken in the same sample with the tasks randomised. All five studies presenting evidence on decision certainty were undertaken in the same sample. However, only two of these studies (Van Dijk et al., and Whitty et al., [52, 57]) randomised the order in which the tasks (DCE or BWS) were seen.

3.6.2 Comparison of DCE and multiprofile-case BWS findings

Conceptually, multiprofile-case BWS is a more similar task to traditional format DCE, than is profile-case BWS. In the only study comparing DCE and multiprofile-case BWS, Xie et al. reported similar preference estimates for both approaches, but wider 95% confidence intervals around the preference estimated for DCE than for multiprofile-case BWS [53]. They observed a smaller variance in estimating latent utilities with BWS than with DCE suggesting greater decision certainty (less randomness in decision-making) for the eight BWS than the twelve DCE tasks, even though the participants saw the same 48 profiles in both methods. Hence, multiprofile-case BWS may need a smaller sample size (for the same level of prediction accuracy and precision). This might be expected, given the BWS tasks used data on both best and worst from triple choice sets; whereas, the DCE used best data from pairs. Therefore, the DCE had less ranking data (12 rank observations from pairs) than the BWS (16 rank observations from triplets).

3.6.3 Comparison of profile and multiprofile-case BWS results

Two studies compared preferences estimated from profile and multiprofile-case BWS tasks. Weernink et al. explored treatment preferences for Parkinson's Disease [54]. They reported highly comparable utility values at the aggregate (Pearson correlation R2 0.98, $p<0.001$) and individual (within

person) analysis level (Pearson correlation R2 0.97, p<0.001). However, there was a difference in ranking of the importance of some attributes. Participants also assigned slightly higher utility values to treatment profiles with multiprofile than profile-case BWS.

In the second study, Yoo & Doiron explored preferences for different nursing jobs [37]. Whilst they found improvements in non-pecuniary job attributes were valued similarly by both methods, participants placed greater value on the pecuniary (salary) gains over non-pecuniary gains in the multiprofile case. Thus, estimates of marginal rates of substitution based on their data would be likely to differ between the elicitation formats.

3.7 Study quality

On average, the included studies achieved a score of 3.42 (out of a maximum of 5) in the PREFS checklist (Supplementary material, Table S1). Studies comparing BWS with DCE achieved a mean score of 3.5 and those comparing different BWS cases achieved a mean score of 3. “Responders” and “Findings” were generally poorly reported. For “Responders”, this was because evidence was lacking on response rates and indicating whether responders were similar to non-responders. Cheung [2] recommends tracking demographic variables to assess possible differences between the two groups. However, the current widespread use of online panels (including in the studies in this review) might hinder the tracking of demographics for non-responders. For “Findings”, this was mainly because studies either did not *explicitly* state that data from all responders who completed or partially completed the preference tasks were included in the analysis, or if some were excluded (e.g. for failing a validity test) did not provide evidence that those excluded did not differ from those included or that the findings did not differ with their inclusion/exclusion.

4. DISCUSSION

DCE and BWS stated preference methods are gaining popularity to evaluate outcomes and priorities in health and health care. Moreover, they are now prominent approaches for health state valuation – which informs access decisions for health care services and technologies across many countries and health jurisdictions [58, 59]. Understanding the merits and limitations of these methods is paramount to the accurate and reliable valuation of health outcomes and healthcare, to support consistent

access decisions. Importantly, DCE and BWS methods are being used instead of (or potentially alongside) each other to derive quantitative weights. Regardless of what these weights are called (e.g. preferences, utilities, or values), the findings of DCE and BWS tasks are being used to derive weights that are intended to measure the same construct and to be interchangeable with each other in predicting choice. Whilst evidence is limited and the findings of this review are discordant, the balance of evidence suggests that the "preferences" derived from DCE and profile-case BWS tasks may not interchangeable. This is particularly the case if the preferences are used to derive marginal rates of substitution, as opposed to a simple summary of preference direction or rank of beneficial attributes at the extremes of the preference space. Most applications of stated preferences in healthcare decision-making, including estimates for health state valuation or risk:benefit assessment, require marginal rates of substitution to be estimated. This review suggests we cannot assume that DCE and profile-case BWS will imply the same decision for these purposes.

Whilst the finding of a lack of concordance between DCE and profile-case BWS is concerning, comparisons of other stated preference methods (such as TTO and standard gamble, SG) also suggest they provide different estimates of preferences used to derive key trade-offs for decision-making in health [60, 61]. Therefore, this review does not necessarily suggest a "new" problem with use of DCE or BWS methods. Rather, it suggests we need to be cautious and understand and explore these differences further, why they occur, how they are influenced, and their implications for decision-making, in order to understand the relative merits and choice between the DCE and BWS methods. Whilst it has previously been postulated that the comparative performance of DCE and BWS methods might be context dependent (with poorer performance in priority-setting contexts) [17], this does not seem to be supported by an assessment of the current evidence (with the possible exception of comparative acceptability, which is discussed below). Some authors have argued that the DCE is a superior approach to BWS because its measurement properties have been extensively examined [38]. However, the measurement properties of an instrument are context dependent, and so the very limited number of studies examining these methods in different health care contexts, perhaps most notably in health state valuation which is a very specific framing of the task, limits any conclusions regarding their comparative merits.

A low level of concordance between methods suggests that either both methods are measuring the same construct but one is doing so more accurately than the other; or that the methods are measuring different constructs. The studies identified in this review do not provide conclusive evidence on which of these is most likely. However, they do provide us with relevant insights. If both methods are measuring the same construct (i.e. preferences) but one is doing so more accurately than the other, this would suggest one method is more valid than the other. The review provided only limited evidence on the comparative internal validity of the DCE and BWS formats. Most evidence suggests either equal internal validity, or favours DCE over profile-case BWS. However, whilst some evidence appears to favour DCE, close examination of the tests on which that evidence is based and of the probabilities of passing those tests by chance alone suggests that generally these tests may be unreliable for comparing the internal validity of DCE and profile-case BWS tasks. There was no evidence found on the convergent validity of DCE as compared to *profile-case* BWS. This should be considered a research priority. Given the increased use of both these formats for health state valuation, it would be helpful to establish which, if either, might give the most similar results to other stated preference methods used in health state valuation (such as TTO), and therefore might be used most consistently alongside other methods in decision-making. We conclude there is currently insufficient comparative evidence available on validity to suggest any superiority of either DCE or BWS methods.

An assessment of external validity would provide insight into whether DCE and BWS measure the same construct. This is challenging, as there is no accepted gold standard measure of preferences and external validity is difficult to assess for stated preference methods in the regulated healthcare environment. There are theoretical reasons to suggest that DCE and BWS, particularly profile-case BWS, may not be measuring the same construct. For example, the single profile format of profile-case BWS does not elicit a trade-off as is the case for conventional format DCE or multiprofile BWS, selecting a 'worst' alternative is a conceptually different task to selecting a 'best' alternative, and the max-diff or sequential decision models assumed for BWS are not required for DCE. Therefore, arguably we should not necessarily expect the findings of a BWS task to be concordant with those of a DCE. Nevertheless, this raises a challenge if the preferences from both methods (and indeed other methods giving dissimilar findings) are used interchangeably in decision-making. Researchers and decision-makers need to be aware that DCE and profile-case BWS may not be measuring the same construct and

therefore may imply different decisions in health if the same decision-making threshold is used. Given the recognised importance of patient and public preferences in healthcare decision-making, consideration needs to be given to how preferences from DCE and BWS can be integrated into economic evaluations and healthcare decisions. Such incorporation would mean that licensing, reimbursement and policy decisions could reflect stakeholders' preferences in a more accurate way, thus contributing to greater efficiency, satisfaction and adherence to health interventions and programmes [34]. For integration to develop further, researchers need to agree on standard methodologies to undertake DCE and BWS stated preference studies and to assess in which circumstances one format is preferred to another.

Evidence from studies that explored the comparative error variance of choice models suggests that profile-case BWS may be more statistically efficient than DCE; although, this may not be the case in priority-setting contexts. Thus, the profile-case BWS may require a smaller sample size for a desired level of precision, making the choice data potentially more economical to collect for a profile-case BWS than for a DCE study. Whilst this supports the assertion that BWS is a more statistically efficient format [62], it does not suggest that BWS has greater response efficiency which is also an important consideration for sample size and data quality [40].

The acceptability of the method for participants provides important information on the likely response efficiency with which a stated preference study can be completed and quality of the data collected. If the two methods are equally valid, the choice for which method to use might be strongly informed by the comparative acceptability of the tasks for participants. The majority of evidence on comparative acceptability of DCE and BWS focussed on the participants' self-reported task difficulty or preferred task. Evidence emerging from the review suggests participants find DCE tasks at least as easy to complete as profile-case BWS tasks, which is contrary to the general inference from the literature, in which some have argued that choosing best and worst attribute levels from a single profile should be a simpler task than choosing a preferred profile from two or more alternative profiles in a DCE (or multiprofile BWS) [20, 62]. We find no evidence in this review to support this assertion. Two studies provided evidence suggesting that participants found the DCE task easier than the profile-case BWS task and in one of these three quarters of participants also indicated they preferred the DCE task; both studies were undertaken in a priority-setting context [17, 44, 51]. Therefore, whilst there is no

evidence to suggest a difference in acceptability in most contexts, it might be the case, that in the context of priority-setting where social preferences are elicited, the DCE performs better that profile-case BWS in terms of acceptability.

Although the number of published DCE and BWS studies has been increasing [1, 2], the number of publications comparing both elicitation procedures or different BWS cases remains low. This review is dependent on published studies. Therefore, there is a risk some relevant studies may not have been identified for inclusion, for example due to publication bias. One of the purported advantages of BWS is that if the task contains more than three attribute levels (profile case) or alternative profiles (multiprofile case), it is possible to obtain a full ranking from each choice set in a sequential fashion (e.g. best, worst, next best, next worst etc), providing more data and possibly supporting quantification of individual level preferences [63]. None of the included studies used repeated rounds of BWS choices; therefore, any potential merit of this approach has not been considered. Additionally, several authors have highlighted the importance of considering interaction effects [1, 64], and this is likely to be particularly important in the health state valuation context. If authors do not consider them but instead implicitly assume that the attribute effects are independent, and this assumption proves to be false, then preference estimates will be biased [65]. Few studies considered interaction effects, and they were only considered for DCE and not BWS. Consideration of interaction effects in profile-case BWS is technically problematic, and any merit of DCE compared to BWS in terms of considering interactions has not been identified in this review. There were few studies that explored the comparative merits of multiprofile-case BWS; yet, the format of multiprofile-case BWS makes it conceptually more similar to a DCE than profile-case BWS, which has been commonly applied. It would be logical to explore the merits of multiprofile BWS in health further. No study explored the comparative merits of object case BWS. However, object-case BWS is conceptually more similar to a simple ranking exercise than to a DCE, which might explain its absence from this review. This review has raised a number of methodological research priorities, which could be addressed with further comparative studies in the future.

5. CONCLUSION

The very limited evidence available suggests that DCE and profile-case BWS may be of equal validity in eliciting preferences in a health context, but give different preference estimates – suggesting they may be measuring different constructs. Therefore, choice between methods is likely to based on normative considerations related to coherence of the methods with the theoretical frameworks for which the preferences are to be used and on pragmatic considerations related to ease of data collection. There is insufficient evidence to suggest the decision context (e.g. health state valuation or priority-setting) impacts the comparative merits of the DCE and BWS tasks, with the possible exception of participant acceptability, which might favour DCE over profile-case BWS when preferences are elicited to inform priority-setting. Profile-case BWS may be a more statistically efficient approach to data collection than DCE, requiring a smaller sample size for a given level of precision. There is insufficient evidence to draw conclusions on the comparative merits of DCE and either object of multiprofile-case BWS. Overall, the limitations in the evidence mean that this summary is indicative rather than conclusive with further data required to expand our understanding of the comparative merits of these methods. There are clear implications for the integration of DCE and BWS methods to inform health technology decisions including through the derivation of utility values and risk:benefit trade-offs. Therefore, expanding our understanding of the comparative merits of these approaches should be a research priority.

6 REFERENCES

1. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health economics. 2012;21(2):145-72. doi:10.1002/hec.1697.

2. Cheung KLW, B. F. M.; Hollin, I. L.; Janssen, E. M.; Bridges, J. F.; Evers, S. M. A. A.; Hiligsmann, M. Using Best-Worst Scaling to Investigate Preferences in Health Care. PharmacoEconomics. 2016:1-15.

3. Whitty JA, Lancsar E, Rixon K, Golenko X, Ratcliffe J. A systematic review of stated preference studies reporting public preferences for healthcare priority setting. The Patient. 2014;7(4):365-86.

4. Whitty JA, Ratcliffe J, Kendall E, Burton P, Wilson A, Littlejohns P et al. Prioritising patients for bariatric surgery: building public preferences from a discrete choice experiment into public policy. BMJ open. 2015;5(10):e008919. doi:10.1136/bmjopen-2015-008919.

5. Marsh K, M IJ, Thokala P, Baltussen R, Boysen M, Kalo Z et al. Multiple Criteria Decision Analysis for Health Care Decision Making--Emerging Good Practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force. Value Health. 2016;19(2):125-37. doi:10.1016/j.jval.2015.12.016.

6. Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z et al. Multiple Criteria Decision Analysis for Health Care Decision Making--An Introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force. Value Health. 2016;19(1):1-13. doi:10.1016/j.jval.2015.12.003.

7. Drummond MF, Sculpher MJ, Torrance GW, O'Brien B, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. 3rd ed. New York: Oxford University Press; 2005.

8. Nord E, Street A, Richardson J, Kuhse H, Singer P. The significance of age and duration of effect in social evaluation of health care. Health Care Anal. 1996;4(2):103-11.

9. Olsen JA, Donaldson C. Helicopters, hearts and hips: using willingness to pay to set priorities for public sector health care programmes. Soc Sci Med. 1998;46(1):1-12.

10. Ryan M, Scott DA, Reeves C, Bate A, van Teijlingen ER, Russell EM et al. Eliciting public preferences for healthcare: a systematic review of techniques. Health Technol Assess. 2001;5(5):1-186.

11. Ryan M, Gerard K. Using discrete choice experiments to value health care programmes: current practice and future research reflections. Applied health economics and health policy. 2003;2(1):55-64.

12. Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob EW. Discrete Choice Experiments in Health Economics: A Review of the Literature. PharmacoEconomics. 2014;32(9):883-902. doi:10.1007/s40273-014-0170-x.

13. Thurstone LL. A law of comparative judgment. Psychological Review. 1994;101(2):266-70. doi:10.1037/0033-295X.101.2.266.

14. McFadden D. Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka P, editor. Frondtiers in Econometrics. New York: Academic Press; 1974.

15. Louviere JJF, T. N. Using best-worst scaling choice experiments to measure public perceptions and preferences for healthcare reform in Australia. Patient. 2010;3(4):275-83.

16. Lancaster KJ. A new approach to consumer theory. J Polit Econ. 1966;74. doi:10.1086/259131.

17. Whitty JAR, J.; Chen, G.; Scuffham, P. A. Australian Public Preferences for the Funding of New Health Technologies: A Comparison of Discrete Choice and Profile Case Best-Worst Scaling Methods. Medical decision making : an international journal of the Society for Medical Decision Making. 2014;34(5):638-54. doi:10.1177/0272989x14526640.

18. Finn A, Louviere JJ. Determining the approporiate response to evidence of public concern: the case of food safety. J Public Policy Marketing. 1992;11(1):12-5.

19. Louviere JJ, Woodworth GG. Best-Worst Scaling: A Model for Largest Difference Judgments.: University of Alberta: Working Paper, Faculty of Business.1990.

20. Marley AAJ, Louviere JJ. Some probabilistic models of best, worst, and best–worst choices. Journal of Mathematical Psychology. 2005;49(6):464-80. doi:http://dx.doi.org/10.1016/j.jmp.2005.05.003.

21. Louviere JJ, Flynn TN, Marley AAJ. Best-worst scaling: Theory, methods and applications. Best-Worst Scaling: Theory, Methods and Applications. 2015.

22. Flynn TN. Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. Expert Rev Pharmacoecon Outcomes Res. 2010;10(3):259-67. doi:10.1586/erp.10.29.

23. Cheung KL, Wijnen BF, Hollin IL, Janssen EM, Bridges JF, Evers SM et al. Using Best-Worst Scaling to Investigate Preferences in Health Care. Pharmacoeconomics. 2016;34(12):1195-209. doi:10.1007/s40273-016-0429-5.

24. Louviere JJ, Flynn TN. Using best-worst scaling choice experiments to measure public perceptions and preferences for healthcare reform in australia. Patient. 2010;3(4):275-83. doi:10.2165/11539660-000000000-00000

01312067-201003040-00008 [pii].

25. Ratcliffe J, Flynn T, Terlich F, Stevens K, Brazier J, Sawyer M. Developing Adolescent-Specific Health State Values for Economic Evaluation: An Application of Profile Case Best-Worst Scaling to the Child Health Utility 9D. Pharmacoeconomics. 2012;30(8):713-27. doi:10.2165/11597900-000000000-00000.

26. Al-Janabi H, Flynn TN, Coast J. Estimation of a Preference-Based Carer Experience Scale. Med Decis Making. 2011;31:458-68. doi:0272989X10381280 [pii]

10.1177/0272989X10381280.

27. Flynn TN, Louviere JJ, Peters TJ, Coast J. Estimating preferences for a dermatology consultation using Best-Worst Scaling: comparison of various methods of analysis. BMC Med Res Methodol. 2008;8:76. doi:1471-2288-8-76 [pii]

10.1186/1471-2288-8-76.

28. Yoo HI, Doiron D. The use of alternative preference elicitation methods in complex discrete choice experiments. J Health Econ. 2013;32(6):1166-79. doi:10.1016/j.jhealeco.2013.09.009.

29. Coast J, Flynn TN, Natarajan L, Sproston K, Lewis J, Louviere JJ et al. Valuing the ICECAP capability index for older people. Soc Sci Med. 2008;67(5):874-82. doi:S0277-9536(08)00254-2 [pii]

10.1016/j.socscimed.2008.05.015.

30. Coast J, Salisbury C, de Berker D, Noble A, Horrocks S, Peters TJ et al. Preferences for aspects of a dermatology consultation. Br J Dermatol. 2006;155(2):387-92.

31. Cameron MP, Newman PA, Roungprakhon S, Scarpa R. The marginal willingness-to-pay for attributes of a hypothetical HIV vaccine. Vaccine. 2013;31(36):3712-7. doi:10.1016/j.vaccine.2013.05.089.

32. Lancsar E, Louviere J, Donaldson C, Currie G, Burgess L. Best worst discrete choice experiments in health: methods and an application. Soc Sci Med. 2013;76(1):74-82. doi:10.1016/j.socscimed.2012.10.007.

33. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ. 2012;21(2):145-72. doi:10.1002/hec.1697.

34. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA et al. Conjoint analysis applications in health--a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2011;14(4):403-13. doi:10.1016/j.jval.2010.11.013.

35. Louviere J, Lings I, Islam T, Gudergan S, Flynn T. An introduction to the application of (case 1) best–worst scaling in marketing research. International Journal of Research in Marketing. 2013;30(3):292-303. doi:http://dx.doi.org/10.1016/j.ijresmar.2012.10.002.

36. Flynn TN, Louviere JJ, Peters TJ, Coast J. Best-worst scaling: What it can do for health care research and how to do it. J Health Econ. 2007;26(1):171-89.

37. Yoo HID, D. The use of alternative preference elicitation methods in complex discrete choice experiments. Journal of Health Economics. 2013;32(6):1166-79.

38. Krucien NW, Verity; Ryan, Mandy. Is Best–Worst Scaling Suitable for Health State Valuation? A Comparison with Discrete Choice Experiments. Health economics. 2016:n/a-n/a. doi:10.1002/hec.3459.

39. Mühlbacher AC, Bethge S. Patients' preferences: a discrete-choice experiment for treatment of non-small-cell lung cancer. The European Journal of Health Economics. 2015;16(6):657-70. doi:10.1007/s10198-014-0622-4.

40. Johnson RF, Lancsar E, Marshall D. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2013;16. doi:10.1016/j.jval.2012.08.2223.

41. Lancsar E, Swait J. Reconceptualising the External Validity of Discrete Choice Experiments. PharmacoEconomics. 2014;32(10):951-65. doi:10.1007/s40273-014-0181-7.

42. Mas-Colell AW, Michael; Green, Jerry. Microeconomic Theory. Oxford University Press; 1995.

43. Joy SM, Little E, Maruthur NM, Purnell TS, Bridges JFP. Patient Preferences for the Treatment of Type 2 Diabetes: A Scoping Review. PharmacoEconomics. 2013;31(10):877-92. doi:10.1007/s40273-013-0089-7.

44. Whitty JAW, R.; Golenko, X.; Ratcliffe, J. A think aloud study comparing the validity and acceptability of discrete choice and best worst scaling methods. PloS one. 2014;9(4):e90635. doi:10.1371/journal.pone.0090635.

45. Netten AB, P.; Malley, J.; Potoglou, D.; Towers, A. M.; Brazier, J.; Flynn, T.; Forder, J.; Wall, B. Outcomes of social care for adults: Developing a preference-weighted measure. Health Technology Assessment. 2012;16(16):1-165.

46. Potoglou DB, Peter; Flynn, Terry; Netten, Ann; Malley, Juliette; Forder, Julien; Brazier, John E. Best–worst scaling vs. discrete choice experiments: An empirical comparison using

social care data. Social Science & Medicine. 2011;72(10):1717-27. doi:10.1016/j.socscimed.2011.03.027.

47. Coast J, Huynh E, Kinghorn P, Flynn T. Complex Valuation: Applying Ideas from the Complex Intervention Framework to Valuation of a New Measure for End-of-Life Care. PharmacoEconomics. 2016;34(5):499-508. doi:10.1007/s40273-015-0365-9.

48. Flynn TJP, Tim; Coast, Joanna. Quantifying response shift or adaptation effects in quality of life by synthesising best-worst scaling and discrete choice data. Journal of Choice Modelling. 2013;6:34-43. doi:http://dx.doi.org/10.1016/j.jocm.2013.04.004.

49. Hollin ILP, H. L.; Bridges, J. F. P. Caregiver Preferences for Emerging Duchenne Muscular Dystrophy Treatments: A Comparison of Best-Worst Scaling and Conjoint Analysis. Patient. 2014;8(1):19-27.

50. Janssen EMS, J. B.; Bridges, J. F. P. A Framework for Instrument Development of a Choice Experiment: An Application to Type 2 Diabetes. Patient. 2016;9(5):465-79.

51. Severin FS, J.; Muhlbacher, A.; Rogowski, W. H. Eliciting preferences for priority setting in genetic testing: A pilot study comparing best-worst scaling and discrete-choice experiments. European Journal of Human Genetics. 2013;21(11):1202-8.

52. van Dijk JDG-O, Catharina G. M.; Marshall, Deborah A.; Ijzerman, Maarten J. An Empirical Comparison of Discrete Choice Experiment and Best-Worst Scaling to Estimate Stakeholders' Risk Tolerance for Hip Replacement Surgery. Value in Health. 2016;19(4):316-22. doi:10.1016/j.jval.2015.12.020.

53. Xie FP, E.; Gaebel, K.; Oppe, M.; Krabbe, P. F. M. Eliciting preferences to the EQ-5D-5L health states: Discrete choice experiment or multiprofile case of best-worst scaling? European Journal of Health Economics. 2014;15(3):281-8.

54. Weernink MGG-O, C. G.; I. Jzerman MJ; van Til, J. A. Valuing Treatments for Parkinson Disease Incorporating Process Utility: Performance of Best-Worst Scaling, Time Trade-Off, and Visual Analogue Scales. Value in health : the journal of the International Society for

Pharmacoeconomics and Outcomes Research. 2016;19(2):226-32. doi:10.1016/j.jval.2015.11.011.

55. Ryan M, Watson V, Entwistle V. Rationalising the 'irrational': a think aloud study of discrete choice experiment responses. Health economics. 2009;18. doi:10.1002/hec.1369.

56. Lancsar E, Louviere J. Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences? Health economics. 2006;15. doi:10.1002/hec.1104.

57. Whitty JA, Ratcliffe J, Chen G, Scuffham PA. Australian Public Preferences for the Funding of New Health Technologies: A Comparison of Discrete Choice and Profile Case Best-Worst Scaling Methods. Med Decis Making. 2014;34:638–54. doi:10.1177/0272989X14526640.

58. Stafinski T, Menon D, Philippon DJ, McCabe C. Health technology funding decision-making processes around the world: the same, yet different. Pharmacoeconomics. 2011;29(6):475-95. doi:2 [pii]

10.2165/11586420-000000000-00000.

59. Yoongthong W, Hu S, Whitty JA, Wibulpolprasert S, Sukantho K, Thienthawee W et al. National drug policies to local formulary decisions in Thailand, china, and australia: drug listing changes and opportunities. Value Health. 2012;15(1 Suppl):S126-31. doi:S1098-3015(11)03540-6 [pii]

10.1016/j.jval.2011.11.003.

60. Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. J Health Econ. 1996;15(2):209-31. doi:0167-6296(95)00038-0 [pii].

61. Arnold D, Girling A, Stevens A, Lilford R. Comparison of direct and indirect methods of estimating health state utilities for resource allocation: review and empirical analysis. BMJ. 2009;339:b2688. doi:10.1136/bmj.b2688.

62. Flynn TNL, J. J.; Peters, T. J.; Coast, J. Best--worst scaling: What it can do for health care research and how to do it. J Health Econ. 2007;26(1):171-89. doi:10.1016/j.jhealeco.2006.04.002.

63. Muhlbacher ACK, A.; Zweifel, P.; Johnson, F. R. Experimental measurement of preferences in health and healthcare using best-worst scaling: an overview. Health Economics Review. 2016;6(1):1-14.

64. Louviere JJ, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Economics, Policy and Law. 2009;4(4):527-46. doi:10.1017/S1744133109990193.

65. Reed Johnson F, Lancsar E, Marshall D, Kilambi V, Mühlbacher A, Regier DA et al. Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value in Health. 2013;16(1):3-13. doi:http://dx.doi.org/10.1016/j.jval.2012.08.2223.

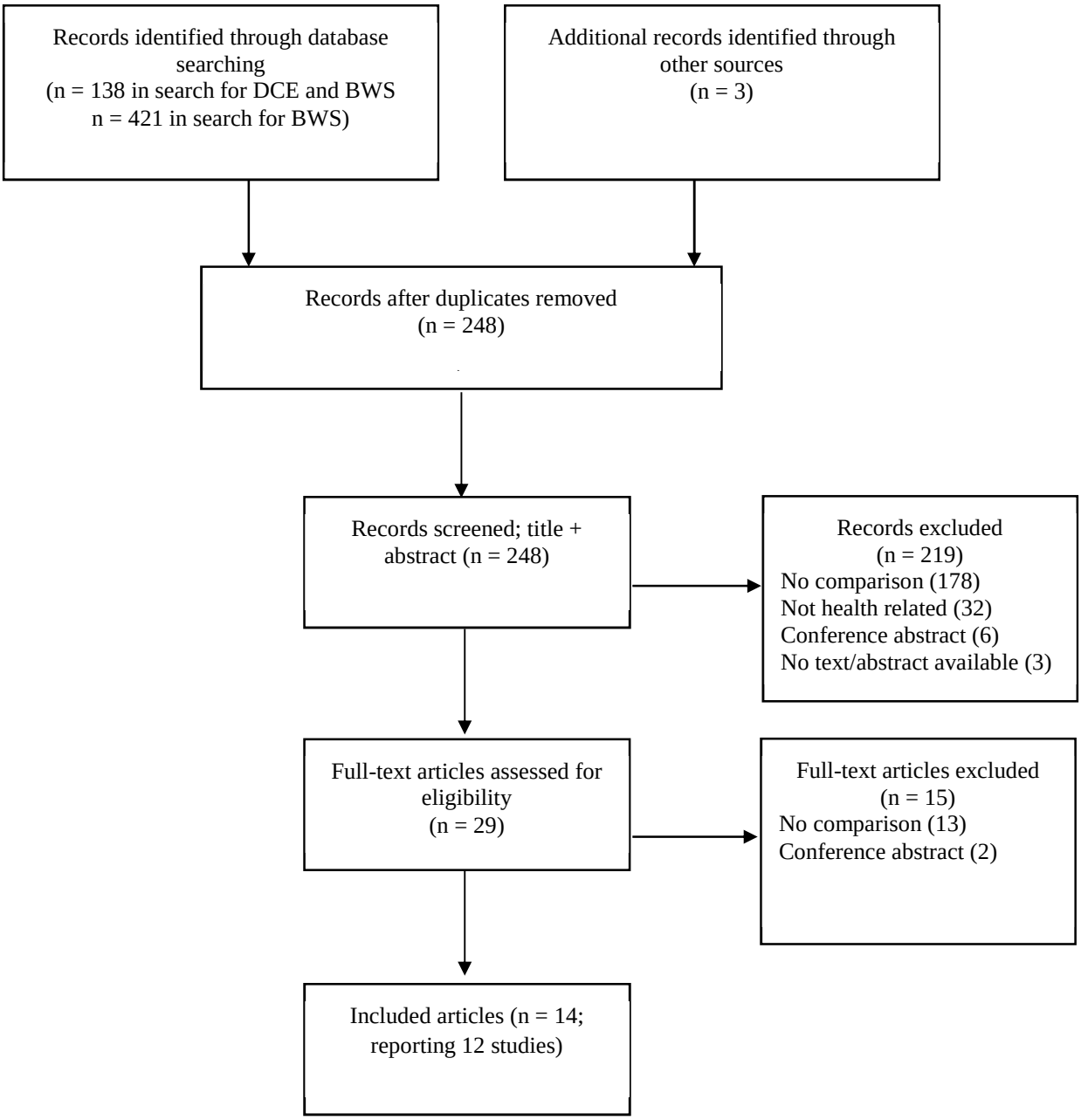

Figure 1 Flow chart of the study identification process

Table 1 Main characteristics of the studies

Study	Country	Target population	Sample size	Application area Perspective	Data available (comparing methods)	Same or different participants for both tasks?
A. Studies comparing DCE with BWS profile case						
Coast et al. [47]	UK	General population (adults)	N=408 across three pilots (Pilots 1,2,3 N=204, 100, 102 respectively)	HSV (end of life) Personal	Concordance	Same (order not randomised)
Flynn [48]	UK	General population (≥65 years)	N= 315	HSV (ICECAP-O) Personal	Concordance	Same (order not randomised)
Hollin [49]	USA	Caregivers	N=119	Treatment options (Duchenne Muscular Dystrophy) For care recipent	Validity Concordance	Same (order not randomised)
Janssen [50]	USA	Patients	Pretest interviews: N=25 Pilot: N=27 (DCE) N=23 (BWS)	Treatment options (type II diabetes mellitus) Personal	Acceptability Concordance	Different
Krucien [38]	UK	Patients	N= 293	HSV (Glaucoma utility Index) Personal	Validity Concordance	Same (order not randomised)

Netten [45], Potoglou [46]	UK	General population (oversample ethnic minorities and ≥65 years)	User interviews: N=30; Pilot N=300	Social care outcome valuation (ASCOT) Personal	Acceptability Validity Concordance	Same (order randomised)
Severin [51]	NL DE	Experts (convenience)	DCE N= 31 BWS: N=26	Priority-setting (genetic testing)	Acceptability Validity Concordance	Different
van Dijk [52]	USA	Potential patients (males, 45-65 years)	N=447 (429 analysed)	Treatment options (hip surgery) Personal	Acceptability Validity Concordance	Same (order randomised)
Whitty [44] (Think aloud study) Whitty [17] (Main study)	AU	General population (adults)	Think Aloud: N=24 Main: N=930	Priority-setting (funding new health technologies) (Social)	Acceptability Validity Concordance	Same (order randomised)
B. Studies comparing DCE with BWS multiprofile case						
Xie [53]	CA	General population (convenience)	N= 100	HSV (EQ-5D-5L) Personal	Acceptability Validity Concordance	Same (order randomised)
C. Studies comparing BWS profile with BWS multiprofile case						
Weernink [54]	UK, NL	General population (18-65 years)	N=613	Treatment options (Parkinsons Disease) Personal	Validity Concordance	Same (not clear if order randomised)
Yoo [37]	AU	Nursing students and graduates	N=526	Job preferences	Concordance	Same (order not randomised)

AU = Australia; CA = Canada; DE= Germany; NL= Netherlands; UK= United Kingdom; USA= United States of America
HSV health state valuation

Table 2 Experimental design and data analysis

Study		Attribute number	Task number[b] (and choice set size)	Statistical design	Analytic method
A. Studies comparing DCE with BWS profile case					
Coast et al. [47]	**DCE**	7	8 (pilots 1,2); 16 (pilot 3) Pairs inc. constant comparator	OMEP 32 profiles	CL (main and 2-way interaction effects)
	BWS	7	8 (pilots 1,2); 16 (pilot 3)	OMEP 32 profiles	CL
Flynn [48]	**DCE**	5	16 Pairs inc. participant-specific comparator	OMEP 16 profiles	CL
	BWS	5	16	OMEP 16 profiles	CL
Hollin [49]	**DCE**	6	18 Single profile, reponse yes/no/don't know	OMEP 18 profiles	Logit
	BWS	6	18	OMEP 18 profiles	CL
Janssen [50]	**DCE**	6	12 Pairs	OMEP 81 sets	CL
	BWS	6	16 or 18 (unclear)	OMEP 16 profiles	CL
Krucien [38]	**DCE**	6	32 Pairs	OMEP foldover 32 profiles	MNL, MXL, GMNL
	BWS	6	32	BWS uses first profile in each DCE pair	MNL, MXL, GMNL

Netten [45], Potoglou [46]	**DCE**	DCE1 [a]: 5 DCE2 [a]: 6	DCE1 [a]: 8 Pairs DCE2 [a]: 8 Pairs	D-efficient, main and 2-way interaction effects (within DCE), 128 sets	MXL (main effects; 2-way interaction effects excluded as not significant)
	BWS	9	12	OMEP 32 profiles	MXL
Severin [51]	**DCE**	6	12 Pairs	D-efficient 12 sets	CL
	BWS	6	12	OMEP 24 profiles	CL
van Dijk [52]	**DCE**	5	8 Triplets	Orthogonal balanced overlap 200 versions of 8 sets	CL
	BWS	5	12	Orthogonal 200 versions of 8 sets	CL
Whitty [44] (Think aloud study) **Whitty [17] (Main study)**	**DCE**	7	6 Pairs	Orthogonal, main and 2-way interaction 72 sets	MNL, SMNL, RPL, GMNL (main and selected 2-way interaction effects)
	BWS	7	6	Uses first profile in each DCE pair	MNL, SMNL, RPL, GMNL
B. Studies comparing DCE with BWS multiprofile case					
Xie [53]	**DCE**	5	24 Pairs	Bayesian D-efficient 48 profiles	Probit
	BWS	5	16 Triplets	Bayesian D-efficient 48 profiles	Ordered logit
C. Studies comparing BWS profile with BWS multiprofile case					
Weernink [54]	**BWS profile**	7	9	D-efficient	MXL, CL
	BWS multiprofile	7	10 Triplets	D-efficient	MXL, CL
Yoo [37]	**BWS profile**	12	8	Uses first profile in each BWS multiprofile triplet	Logit and latent class (max-diff)

	BWS multiprofile	12	8 Triplets	D-optimal Resolution 3 16 sets	Latent class heteroskedastic ROL

CL conditional logit; MNL multinomial logit; MXL mixed logit; GMNL generalised multinomial logit; Max-diff Maximum difference logit; OMEP orthogonal main effects plan; ROL rank ordered logit

[a] Authors split nine attributes across two DCEs, with overlap of some of them

[b] Task number from design, excluding any repeat tasks used for rationaility tests

SUPPLEMENTARY MATERIAL

A. Search strategy: Studies comparing DCE with BWS methods

Database: AMED (Allied and Complementary Medicine) + Embase + MEDLINE(R) via Ovid

Date: 21 October 2016

Search term(s):

discrete choice experiment.ab,ti. AND best worst scaling.ab,ti.
DCE.ab,ti. AND BWS.ab,ti.
discrete choice model.ab,ti. AND best worst scaling.ab,ti.
discrete choice modelling.ab,ti. AND best worst scaling.ab,ti.
conjoint analysis.ab,ti. AND best worst scaling.ab,ti.
conjoint analysis.ab,ti. AND BWS.ab,ti.
conjoint study.ab,ti. AND best worst scaling.ab,ti.
conjoint choice experiment.ab,ti. AND best worst scaling.ab,ti.
stated preference.ab,ti. AND best worst scaling.ab,ti.
parwise choices.ab,ti. AND best worst scaling.ab,ti.
paired comparisons.ab,ti. AND best worst scaling.ab,ti.
discrete choice experiment.ab,ti. AND maximum difference.ab,ti.
discrete choice model.ab,ti. AND maximum difference.ab,ti.
discrete choice modelling.ab,ti. AND maximum difference.ab,ti.
conjoint analysis.ab,ti. AND maximum difference.ab,ti.
conjoint study.ab,ti. AND maximum difference.ab,ti.
conjoint choice experiment.ab,ti. AND maximum difference.ab,ti.
stated preference.ab,ti. AND maximum difference.ab,ti.
parwise choices.ab,ti. AND maximum difference.ab,ti.
paired comparisons.ab,ti. AND maximum difference.ab,ti.

Database: EconLit

Date: 21 October 2016

Search term(s):

AB discrete choice experiment AND AB best-worst scaling
AB DCE AND AB BWS
AB discrete choice model AND AB best worst scaling
AB discrete choice modelling AND AB best worst scaling
AB conjoint analysis AND AB best worst scaling
AB conjoint study AND AB best worst scaling
AB conjoint choice experiment AND AB best worst scaling
AB parwise choices AND AB best worst scaling
AB paired comparisons AND AB best worst scaling
AB discrete choice experiment AND AB maximum difference
AB discrete choice model AND AB maximum difference
AB discrete choice modelling AND AB maximum difference
AB conjoint analysis AND AB maximum difference
AB conjoint study AND AB maximum difference
AB conjoint choice experiment AND AB maximum difference
AB parwise choices AND AB maximum difference
AB paired comparisons AND AB maximum difference

Database: Cochrane Library

Date: 21 October 2016

Search term(s):

#1 "discrete choice experiment":ti or "discrete choice experiment":ab (Word variations have been searched); #2 Best worst scaling:ti or Best worst scaling:ab (Word variations have been searched); #3 #1 and #2
#1 "DCE":ti or "DCE" (Word variations have been searched); #2 "BWS":ti or "BWS":ab (Word variations have been searched); #3 #1 and #2
#1 "discrete choice model":ab or "discrete choice model":ti (Word variations have been searched); #2 "Best worst scaling":ab or "Best worst scaling":ti (Word variations have been searched); #3 #1 AND #2
#1 "discrete choice modelling":ab or "discrete choice modelling":ti (Word variations have been searched); #2 "Best worst scaling":ab or "Best worst scaling":ti (Word variations have been searched); #3 #1 AND #2
#1 "conjoint analysis":ab or "conjoint analysis":ti; #2 "Best worst scaling":ab or "Best worst scaling":ti (Word variations have been searched); #3 #1 AND #2
#1 "conjoint study":ab or "conjoint study":ti (Word variations have been searched); #2 "Best worst scaling":ab or "Best worst scaling":ti (Word variations have been searched); #3 #1 AND #2
#1 "conjoint choice experiment":ab or "conjoint choice experiment":ti (Word variations have been searched); #2 "Best worst scaling":ab or "Best worst scaling":ti (Word variations have been searched); #3 #1 AND #2
#1 "stated preference":ab or "stated preference":ti (Word variations have been searched); #2 "Best worst scaling":ab or "Best worst scaling":ti (Word variations have been searched); #3 #1 AND #2
#1 "parwise choices":ab or "parwise choices":ti (Word variations have been searched); #2 "Best worst scaling":ab or "Best worst scaling":ti (Word variations have been searched); #3 #1 AND #2
#1 "paired comparisons":ab or "paired comparisons":ti (Word variations have been searched); #2 "Best worst scaling":ab or "Best worst scaling":ti (Word variations have been searched); #3 #1 AND #2

#1 "discrete choice experiment":ti or "discrete choice experiment":ab (Word variations have been searched); #2 "maximum difference":ti or "maximum difference":ab (Word variations have been searched); #3 #1 and #2
#1 "dce":ti or "dce":ab; #2 "maximum difference":ti or "maximum difference":ab; #3 #1 and #2
#1 "discrete choice model":ti or "discrete choice model":ab; #2 "maximum difference":ti or "maximum difference":ab; #3 #1 and #2
#1 "discrete choice modelling":ti or "discrete choice modelling":ab; #2 "maximum difference":ti or "maximum difference":ab; #3 #1 and #2
#1 "conjoint analysis":ti or "conjoint analysis":ab; #2 "maximum difference":ti or "maximum difference":ab; #3 #1 and #2
#1 "conjoint study":ti or "conjoint study":ab; #2 "maximum difference":ti or "maximum difference":ab; #3 #1 and #2
#1 "conjoint choice experiment":ti or "conjoint choice experiment":ab; #2 "maximum difference":ti or "maximum difference":ab; #3 #1 and #2
#1 "stated preference":ti or "stated preference":ab; #2 "maximum difference":ti or "maximum difference":ab; #3 #1 and #2
#1 "parwise choices":ti or "parwise choices":ab; #2 "maximum difference":ti or "maximum difference":ab; #3 #1 and #2
#1 "paired comparisons":ti or "paired comparisons":ab; #2 "maximum difference":ti or "maximum difference":ab; #3 #1 and #2

Database: CINAHL

Date: 21 October 2016

Search term(s):

AB discrete choice experiment AND AB best-worst scaling
AB DCE AND AB BWS

AB discrete choice model AND AB best worst scaling
AB discrete choice modelLING AND AB best worst scaling
AB conjoint analysis AND AB best worst scaling
AB conjoint study AND AB best worst scaling
AB conjoint choice experiment AND AB best worst scaling
AB parwise choices AND AB best worst scaling
AB paired comparisons AND AB best worst scaling
AB discrete choice experiment AND AB maximum difference
AB discrete choice model AND AB maximum difference
AB discrete choice modelling AND AB maximum difference
AB conjoint analysis AND AB maximum difference
AB conjoint study AND AB maximum difference
AB conjoint choice experiment AND AB maximum difference
AB parwise choices AND AB maximum difference
AB paired comparisons AND AB maximum difference

Database: PubMed

Date: 21 October 2016

Search term(s):

(discrete choice experiment[Title/Abstract]) AND best worst scaling[Title/Abstract]
DCE[Title/Abstract] AND BWS[Title/Abstract]
(discrete choice model[Title/Abstract]) AND best worst scaling[Title/Abstract]
(discrete choice modelling[Title/Abstract]) AND best worst scaling[Title/Abstract]
conjoint analysis[Title/Abstract] AND best worst scaling[Title/Abstract]
conjoint study[Title/Abstract] AND best worst scaling[Title/Abstract]

conjoint choice experiment[Title/Abstract] AND best worst scaling[Title/Abstract]
pairwise choices[Title/Abstract] AND best worst scaling[Title/Abstract]
paired comparisons[Title/Abstract] AND best worst scaling[Title/Abstract]
discrete choice experiment[Title/Abstract] AND maximum difference[Title/Abstract]
discrete choice model[Title/Abstract] AND maximum difference[Title/Abstract]
discrete choice modelling[Title/Abstract] AND maximum difference[Title/Abstract]
conjoint analysis[Title/Abstract] AND maximum difference[Title/Abstract]
conjoint study[Title/Abstract] AND maximum difference[Title/Abstract]
conjoint choice experiment[Title/Abstract] AND maximum difference[Title/Abstract]
(pairwise[All Fields] AND choices[Title/Abstract]) AND maximum difference[Title/Abstract]
(paired comparisons [Title/Abstract]) AND maximum difference[Title/Abstract]

B. Search strategy: Studies comparing different BWS cases

Database: AMED (Allied and Complementary Medicine) + Embase + MEDLINE(R) via Ovid

Date: 2 December 2016

Search term(s):

((((("best worst" and scaling).ab. or "best worst".af.) and analysis.ab.) or "best worst".ab.) and survey.ab.
maxdiff
(((((((((maximum difference and scaling) or maximum difference) and analysis) or maximum difference) and scoring) or maximum difference) and attribute) or maximum difference) and survey).ab.

Database: EconLit

Date: 2 December 2016

Search term(s):

AB "best worst" AND AB scaling OR AB "best worst" AND AB analysis OR AB "best worst" AND AB survey
AB maxdiff
AB "maximum difference" AND AB scaling OR AB "maximum difference" AND AB analysis OR AB "maximum difference" AND AB scoring OR AB "maximum difference" AND AB attribute OR AB "maximum difference" AND AB survey

Database: Cochrane Library

Date: 21 October 2016

Search term(s):

#1 "Best worst":ti or "Best worst":ab; #2 "scaling":ti or "scaling":ab; #3 "analysis":ti or "analysis":ab; #4 "survey":ti or "survey":ab; #5 "#1 and #2"; #6 "#1 and #3"; #7 "#1 and #4"
AB maxdiff:ab
"maximum difference":ab

Database: CINAHL

Date: 2 December 2016

Search term(s):

AB "maximum difference" AND AB scaling OR AB "maximum difference" AND AB analysis OR AB "maximum difference" AND AB scoring OR AB "maximum difference" AND AB attribute OR AB "maximum difference" AND AB survey
AB maxdiff

AB "best worst" AND AB scaling OR AB "best worst" AND AB analysis OR AB "best worst" AND AB survey

Database: PubMed

Date: 2 December 2016

Search term(s):

(((((("best worst"[Title/Abstract]) AND scaling[Title/Abstract]) OR "best worst"[Title/Abstract]) AND analysis[Title/Abstract]) OR "best worst"[Title/Abstract]) AND survey[Title/Abstract]
maxdiff[Title/Abstract])
(((((((((("maximum difference"[Title/Abstract]) AND scaling[Title/Abstract]) OR "maximum difference"[Title/Abstract]) AND analysis[Title/Abstract]) OR "maximum difference"[Title/Abstract]) AND scoring[Title/Abstract]) OR "maximum difference"[Title/Abstract]) AND attribute[Title/Abstract]) OR "maximum difference") AND survey

C. Study quality

Supplementary Table S1 Evaluation of Study Quality (using PREFS criteria)

	Quality					
Study	P	R	E	F	S	∑
A. Studies comparing DCE with BWS profile case						
Coast et al. [47]	1	0	1	0	1	3
Flynn et al.	1	0	1	0	1	3
Hollin [49]	1	0	1	0	1	3
Janssen [50]	1	1	1	1	1	5
Krucien [38]	1	0	1	0	1	3
Netten [45] Potoglou [46]	1	0	1	0	1	3
Severin [51]	1	1	1	1	1	5
van Dijk [52]	1	0	1	0	1	3
Whitty [17]	1	0	1	1	1	4
B. Studies comparing DCE with BWS multiprofile case						
Xie [53]	1	0	1	0	1	3
C. Studies comparing BWS profile with BWS multiprofile case						
Weernink [54]	1	0	1	0	1	3
Yoo [37]	1	0	1	0	1	3

Coding: ‘1’ reporting and ‘0’ not reporting relevant data.