



Best–Worst Scaling and the Prioritization of Objects in Health: A Systematic Review

Ilene L. Hollin¹ · Jonathan Paskett² · Anne L. R. Schuster² · Norah L. Crossnohere² · John F. P. Bridges²

Accepted: 12 June 2022 / Published online: 15 July 2022
© The Author(s) 2022

Abstract

Background and Objective Best–worst scaling is a theory-driven method that can be used to prioritize objects in health. We sought to characterize all studies of best–worst scaling to prioritize objects in health, to assess trends of using best–worst scaling in prioritization over time, and to assess the relationship between a legacy measure of quality (PREFS) and a novel assessment of subjective quality and policy relevance.

Methods A systematic review identified studies published through to the end of 2021 that applied best–worst scaling to study priorities in health (PROSPERO CRD42020209745), updating a prior review published in 2016. The PubMed, EBSCOhost, Embase, Scopus, APA PsychInfo, Web of Science, and Google Scholar databases were used and were supplemented by a hand search. Data describing the application, development, design, administration/analysis, quality, and policy relevance were summarized and we tested for trends by comparing articles before and after 1 January, 2017. Multivariate statistics were then used to assess the relationships between PREFS, subjective quality, policy relevance, and other possible indicators.

Results From a total of 2826 unique papers identified, 165 best–worst scaling studies were included in this review. Applications of best–worst scaling to study priorities in health have continued to grow ($p < 0.01$) and are now used in all regions of the world, most often to study the priorities of patients/consumers (67%). Several key trends can be observed over time: increased use of pretesting ($p < 0.05$); increased use of online administration ($p < 0.01$), and decreased use of paper self-administered surveys ($p = 0.02$); increased use of heterogeneity analysis ($p = 0.02$); an increase in having a clearly stated purpose ($p < 0.01$); and a decrease in comparing respondents to non-respondents ($p = 0.01$). The average sample size has more than doubled, from 228 to 472 respondents, but formal sample size justifications remain low (5.3%) and unchanged over time ($p = 0.68$). While the average PREFS score remained unchanged at 3.1/5, both subjective quality and policy relevance trended up, but changes were not statistically significant ($p = 0.06$ and $p = 0.13$). Most of the variation in subjective quality was driven by PREFS ($R^2 = 0.42$), but it was also positively associated with policy relevance, heterogeneity analysis, and using a balanced incomplete block design, and was negatively associated with not using developmental methods and an increasing sample size.

Conclusions Using best–worst scaling to prioritize objects is now commonly used around the world to assess the priorities of patients and other stakeholders in health. Best practices are clearly emerging for best–worst scaling. Although legacy measures (PREFS) to measure study quality are reasonable, there may need to be new tools to assess both study quality and policy relevance.

1 Introduction

Many decision makers in health are interested in understanding how patients and other stakeholders prioritize various matters [1, 2]. Priorities refers to topics that are regarded with greater relative importance, and understanding priorities, or what matters most, is essential when time and resources are limited. For example, policymakers may be interested in identifying what questions or concerns are most important to a certain group [3], which outcomes people value the most [4], what factors may be considered most

✉ John F. P. Bridges
John.Bridges@osumc.edu

¹ Department of Health Services Administration and Policy, Temple University College of Public Health, Philadelphia, PA, USA

² Department of Biomedical Informatics, The Ohio State University College of Medicine, Columbus, OH, USA

Key Points

Best–worst scaling is a theory-driven method increasingly being used in health. While best–worst scaling can be applied to study preferences when applied to single or multiple product profiles defined by attributes and levels, it can also be applied to study how a finite set of objects should be prioritized without the use of levels. In this instance, best–worst scaling may be referred to as case 1, object case, object scaling, MaxDiff, or simply as best–worst scaling.

The average number of best–worst scaling studies focusing on prioritization has jumped from under five per year prior to 2017 to now 26.4 per year. It is now also used in all regions of the world and for a wide variety of purposes. The average sample size for best–worst scaling has increased over time, likely owing to the growing use of online panels to sample respondents, and the increase in the likelihood that the study is relevant to policy makers.

While the PREFS measure of study quality has received some criticism in recent years, we find that it is highly associated with a global assessment of subjective quality. This said, we also find that several other factors, including policy relevance and issues associated with both the design and analysis, will impact quality and could be included in future measures.

harmful (or most beneficial) when considering certain outcomes [5], the issues driving decision making [6], or what people might view as having potential future impacts [7]. Prioritization and other types of values elicitation also play a central role in shared decision making [8], can guide regulatory decisions [9], inform patient-centered outcomes research [10], and promote patient-centered care [11, 12].

Various methods exist to measure how patients and other stakeholders prioritize matters in health. These methods traditionally included rating or ranking approaches [13], self-explicated methods [14], and a 2^K conjoint analysis (where K objects are segmented into repeated and paired mutually exclusive and exhaustive subsets using a main-effects orthogonal array as part of a choice experiment) [1, 15]. Increasingly, best–worst scaling (BWS) has attracted more attention as a theory-driven method for the prioritization of objects [16, 17]. Best–worst scaling can be used in a number of choice formats for various purposes, including prioritizing objects. Therefore, the term “objects” is often associated with the application of BWS for the purpose of prioritization. It is also known as MaxDiff, object scaling, BWS case 1, or BWS object case. Hereafter, we refer to this

methodology for prioritization as “BWS” and we refer to the general method encompassing all BWS formats with the use of the parenthetical phrase, “all types.”

We sought to characterize all studies of BWS specifically for the prioritization of objects in health. Given that this method received some attention from a review of BWS (all types), published in 2016 [18], we sought to test for changes in practices for application, development, design, administration/analysis, quality, and policy relevance over time by comparing studies published before and after 1 January, 2017. The review was guided by similar reviews for discrete choice experiments (DCE) [19–21], conjoint analysis [22, 23], and BWS (all types) [24], but we also made efforts to advance this review in several ways. First, we specifically focus on BWS for prioritization and second, we abstract several new domains of interest that contribute to the broader stated preference literature. Our study provides important evidence to researchers, reviewers, readers, and funders who are interested in the study of priorities in health and we aimed to highlight the importance of prioritization methods in guiding medical decision making and health policy.

1.1 Using BWS for Prioritizing Objects

First proposed in the early 1990s [25], BWS is a prioritization method that is grounded in random utility theory [26]. While random utility theory was first conceived for pairwise comparisons, such comparisons can become overly burdensome. For example, if we wanted to prioritize only seven objects, a respondent would need to evaluate 21 choice tasks. Best–worst scaling involves sets of objects larger than two, indicating both best and worst (or most/least) options with regard to some criteria. Econometric or statistical methods can then be used to assign values based on these choices [27]. Best–worst scaling has a number of advantages over pairwise voting. First, applying experimental designs, such as a balanced incomplete block design (BIBD) [28] can allow for a comparison of multiple objects more efficiently. For example, seven objects (say a–g) can be compared in sets of four by applying a BIBD in seven tasks containing four objects each (acge, fgbc, ebafe, gefd, dfca, cdeb, badg). The benefit of this design is that each task is the same size, each object is shown the same number of times (here four times) and with each other object the same number of times (here two times). Second, BWS allows two answers per set (‘best’ and ‘worst’), which provides a more efficient presentation and collection of information.

Priorities can also be indirectly derived from methods that measure preferences, such as DCEs and conjoint analyses [29, 30], or as part of multiple criteria decision analysis approaches, such as swing weighting and the analytical hierarchy process [31]. Preference and multiple criteria decision analysis approaches, however, have their limitations

and biases when they are used to measure priorities, especially because they are not measuring priorities directly and, hence, interact attribute importance with level-difference importance (i.e., the level of importance given to an attribute is a function of the differences in the attribute levels used in the preference or multiple criteria decision analysis study). Best–worst scaling too can have this problem when used to study preferences via the evaluation of single or multiple profiles (what is sometimes referred to as BWS case 2 and case 3). This is not true for the object case of BWS (i.e., case 1) as the objects presented in the choice tasks are not given explicit levels.

There are multiple reasons why the prioritization of objects in health is important to health researchers. Researchers are often presented with a list of objects (e.g., from qualitative studies, Delphi process, or review of endpoints) and additional quantification would be beneficial. On other occasions, there are too many concepts being reviewed for a preference study or some attributes may be difficult to assign levels (e.g., stigma), and a prioritization study may be preferable or a first step towards a preference study [32]. Finally, there are many research questions or study purposes that are better suited to a prioritization task, including the selection of endpoints, optimizing communication in healthcare, assessment of the barriers and facilitators to implementation, and many more.

2 Methods

We conducted a systematic review to explore the use of BWS specifically focused on the prioritization of objects. The review protocol was published by PROSPERO (CRD42020209745) and followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The review was guided by two key aims: (i) to characterize all studies of BWS used for prioritization in health, including documenting changes in the application, development, design, and analysis and (ii) to evaluate quality and policy relevance in these applications.

2.1 Search Strategy

A previous review of BWS (all types) studies was used to identify relevant health applications of BWS published prior to 2016 [24]. The search strategy to identify studies published from 2016 to 2021 used various search terms to describe the BWS method, such as ‘best worst scaling,’ ‘BWS,’ ‘best–worst-scaling,’ ‘best worst,’ ‘best–worst,’ ‘object case,’ ‘maxdiff,’ ‘max diff,’ or ‘maximum difference’. The prior review included all cases (i.e., BWS for measuring both priorities and preferences) and searched only two databases (PubMed and Embase). The updated search strategy

narrowed in on prioritization using BWS (i.e., object case) and expanded the search to include seven databases: PubMed, CINAHL/EconLit/MEDLINE (searched in tandem through EBSCOhost), Embase, Scopus, APA PsycInfo, Web of Science, and Google Scholar. In an effort to not replicate previous work, we limited our search to articles published after 1 January, 2016 [24].

Databases were last searched on 1 February, 2022. This date was selected because the time lag between the cut-off date for inclusion and search date allows databases to update. The final search strategy is presented in Table 1 of the Electronic Supplementary Material. This search was supplemented with hand searches of the reference lists of all included studies and articles that cited *Best–Worst Scaling: Theory, Methods and Applications*, an authoritative textbook on the method [33].

2.2 Inclusion and Exclusion Criteria

Studies were eligible for inclusion if they: (i) used the BWS method for prioritization; (ii) reported primary data collection and analysis; (iii) were published one or after 1 January, 2016 and on or prior to 31 December, 2021; (iv) were available in English; (v) were full-text documents; and (vi) were health related. Studies were excluded if they (i) only used simulated data, (ii) did not have defined population cohorts (e.g., no description of sampling frame), (iii) used another BWS case (i.e., profile case or multi-profile case), (iv) were abstracts only, conference proceedings, commentaries, reviews, or editorials, (v) used purely qualitative methods, (vii) were a secondary analysis of existing data, or (viii) were not health related. In cases where two papers analyzing the same dataset existed, we selected the paper published first and excluded any following papers on the basis of them being secondary data analyses. In cases where one paper analyzed multiple experiments, we selected the experiment with the larger sample size. To increase the fidelity of our search, no limitations were placed on the domain of study during our search or extraction process; the review also identified BWS studies in other fields (e.g., business, environment) for use in another review of papers across domains. Thus, post-hoc analysis and consensus-based sorting were conducted to identify which studies were to be included as a health study or excluded as a non-health study and this health-related inclusion criterion was applied as a final step.

2.3 Study Screening

We rescreened studies identified by the prior review because our method of counting studies differed. The prior review focused on counting actual BWS experiments and some papers may have reported on more than one experiment. In this review, we included only one BWS experiment per

paper. In cases in which a single paper included multiple BWS experiments, consistent with our inclusion criteria, only data from the experiment with the largest sample size were included. Similarly, we included only one paper per experiment. In cases in which there were multiple papers published from a single dataset, we include only the first paper, as the second paper would be excluded on the basis of being a secondary analysis of existing data.

We uploaded all studies to Covidence systematic review software (Veritas Health Innovation, Melbourne, VIC, Australia) for review. Duplicate studies were identified by Covidence and by manual review on the basis of title, author, and year. The review for inclusion was conducted in three stages: title review, abstract review, and full-text review. The title review first followed by the abstract review has been shown to be a more efficient approach than screening titles and abstracts together and yields the same number of eligible studies [34].

During the title review, two of four reviewers (IH, NC, SH, JB) independently assessed studies on the basis of title alone and voted accordingly. If there was not enough

information to exclude on title alone, the study moved to the next review phase. The second review stage focused on abstract review. Two of four reviewers (IH, NC, SH, JB) independently assessed studies for inclusion and voted accordingly. Conflicts were resolved by group consensus. Before proceeding to the next review stage, an additional reviewer (JP) performed a quality check of all excluded articles to ensure no articles had been incorrectly excluded.

During the final review stage, two of five reviewers (IH, NC, SH, JP, JB) independently assessed full texts to determine eligibility and voted accordingly. Each full text was reviewed by two different reviewers. Conflicts were resolved by group consensus. After a full-text review, a quality check was performed again on the excluded articles to confirm no studies were misclassified. Checks were also performed looking for references that used the same data across multiple articles. In these cases, the studies were grouped together, and the earliest published article was used for data extraction. After extraction, studies were classified by domain and non-health domains were excluded from this analysis.

Table 1 Changes in the application of best-worst scaling in health over time

	2010–2016 <i>n</i> = 34	2017–2021 <i>n</i> = 131	<i>P</i> value
Number of studies per year, mean [range]	4.9 [1–15]	26.4 [11–37]	< 0.01
Region, <i>n</i> (%)			0.94
North America	16 (47.1%)	55 (42.0%)	
Europe	8 (23.5%)	28 (21.2%)	
Asia	3 (8.8%)	14 (10.7%)	
Africa	1 (2.9%)	7 (5.3%)	
Oceania	2 (5.9%)	6 (4.6%)	
South America	0 (0.0%)	2 (1.5%)	
Multi-continent	4 (11.8%)	15 (11.5%)	
Not specified	0 (0.0%)	4 (3.1%)	
Terminology used to describe, <i>n</i> (%)			0.04
BWS	23 (67.6%)	89 (67.9%)	
BWS object case	0 (0.0%)	17 (13.0%)	
MaxDiff	7 (20.6%)	12 (9.2%)	
BWS case 1	3 (8.8%)	11 (8.4%)	
Best worst choice/task	0 (0.0%)	2 (1.5%)	
Object scaling	1 (2.9%)	0 (0.0%)	
Study objective*, <i>n</i> (%)			
Substantive/empirical	33 (97.1%)	128 (97.7%)	0.83
Educational/methodological	3 (8.8%)	2 (1.5%)	0.03
Pilot	1 (2.9%)	2 (1.5%)	0.58
Perspective*, <i>n</i> (%)			
Patient/consumer	20 (58.8%)	88 (66.7%)	0.39
Provider/producer	13 (38.2%)	43 (32.8%)	0.55
Citizen/societal	2 (5.9%)	10 (7.6%)	0.73

BWS best–worst scaling

*All that apply

2.4 Data Extraction

We developed an extraction form based on the general framework for carrying out a BWS study wherein the categories of information collected represent stages of decision making in the development and conduct of a BWS. Individual extraction items within each category were adapted from those used in previous stated-preference reviews [24, 35], and selected based on information relevant to our review objectives. The extraction form was designed collaboratively during group discussion. Four reviewers (IH, NC, JP, JB) piloted the extraction form in a total of nine studies, leading to clarification and consolidation of extraction fields.

The final form included space to extract data across five categories: application type, development, design, administration/analysis, and evaluation. Application-type items included first author, journal title, year of publication, literature (i.e., domain), country (recategorized as region), terminology used to describe the BWS method, study objective, and perspective (i.e., whose preferences are measured). Development items included the ways in which objects were developed for the survey (e.g., literature review, formal qualitative methods). Design items included mode of survey administration (e.g., online), time frame of prioritization scenario (i.e., past, present, or future), measurement scale, experimental design, BWS anchor description (i.e., most/least, best/worst), total number of objects, number of objects per task, number of tasks in the experiment, and number of tasks per respondent. Administration/analysis items included sample size and its justification, analytic program used, statistical analyses, theoretical assumptions, directionality of results, and type of heterogeneity analysis (if applicable). Evaluation items included two assessments of study quality and policy relevance. Information from included studies was extracted by a single reviewer (IH, JP, NC, AS, or JB).

2.5 Study Quality and Policy Relevance

We used the PREFS checklist, a five-point checklist for assessing the quality and risk of bias of preference studies [36]. Developed as part of a scoping review on patient preferences in diabetes mellitus, it has since been used in reviews of preference literature [37–40]. The PREFS checklist evaluates reporting according to five criteria including: purpose of study, respondent sampling, explanation of assessment methods, complete sets of responses included in the findings, and use of significance testing. Scores range from zero to five. Limitations of PREFS have been previously noted [41], such as that it does not include several important criteria for assessing the quality of health-related stated preference studies, such as heterogeneity assessment, sample size, or patient centrality, and that it may lack specificity to discriminate across studies.

Given the limitations of PREFS, we included a global, single-item, subjective quality assessment. Global assessment scales of quality are routinely included as a part of the quality review or critical appraisal process for measurement instruments [42]. Subjective quality scores were awarded on a scale of 1–10 at each reviewer's discretion based on their impression of the overall quality of the study.

A global assessment of policy relevance was used to assess potential policy relevance [36]. These policy relevance scores are subjective ratings of policy relevance that were awarded on a scale of 1–10 at each reviewer's discretion. Policy relevance scores focused on the perceived utility of the study to policy and decision makers. The studies were scored by a single reviewer (IH, JP, NC, AS, or JB). Both the subjective quality scores and policy relevance scores were intended to be exploratory subjective indicators that could be used to assess the strengths and weaknesses of the PREFS measure, in addition to assessing differences over time. A full validation of these measures was not intended as part of this study and beyond the scope of this paper. This said, they did prove to be beneficial and statistically meaningful in subsequent analyses of the PREFS measure.

We explored the association between PREFS and the subjective quality score and other variables, including policy relevance. All associations were explored using ordinary least squares on the full data set of all studies in both time periods and robust standard errors are reported. Primary analyses assessed to what extent PREFS was associated with subjective quality. The independent variable of these models was subjective quality and the dependent variable was PREFS (as a total score or as five individual criteria). We hypothesized that PREFS would be associated with subjective quality.

We also ran additional analyses exploring the association between subjective quality and variables that have been hypothesized to reflect study quality including: policy relevance [43], heterogeneity analysis [44], the absence of developmental methods [45], use of a BIBD [28], and sample size [46]. The purpose of these analyses was to identify gaps in the PREFS checklist that could inform the development of future standards [47]. We hypothesized that these variables would independently predict subjective quality even after accounting for PREFS. Additional exploratory analyses were run exploring associations with policy relevance or PREFS. We hypothesized that policy relevance would not be associated with PREFS. We also hypothesized that variables associated with subjective quality or PREFS would not be associated with policy relevance.

2.6 Synthesis and Analysis Methods

Studies were grouped based on year of publication: those published prior to 2017 and those from 2017 through 2021.

Abstracted characteristics of studies were compared across these two time periods using statistical hypothesis testing. Chi-squared tests were used to test for differences in proportions of study characteristics that were categorical (e.g., region, study objective) across the two time periods, including when categories were not mutually exclusive. Two-sample t-tests were used to test for differences in the means of characteristics that were continuous (e.g., sample size, number of tasks).

3 Results

3.1 Studies Identified

We screened 26 BWS studies published between 2010 and 2015 identified by the prior review. Seven of these studies were excluded from our analysis. One was reclassified by our research team as ‘not BWS object case,’ three were published in 2016 and included in our updated search, and three were from papers analyzing the same dataset, in which case, consistent with our inclusion criteria, only data from the primary paper were included.

For studies published between 2016 and 2021, the database search strategy identified 5538 studies (PubMed: 1258, CINAHL/EconLit/MEDLINE via EBSCOhost: 894, Embase: 710, Scopus: 1085, APA PsycInfo: 246, Web of Science: 927, Google Scholar: 348). Seventy studies were identified through other sources such as hand searching references and the previous review. After removing duplicates, 2826 unique studies remained.

During the title review stage, 2031 studies were excluded. There were 369 conflicts, all of which were adjudicated by group consensus. Aggregate concordance was high, with a percent agreement of 87% and a kappa statistic of 0.68, indicating a substantial level of agreement between reviewers [48, 49]. During the abstract review stage, we assessed 795 studies and excluded 245 studies. There were 157 conflicts, producing a percent agreement of 80% and kappa statistic of 0.53.

A full-text review was conducted on 550 studies. A full-text review produced a percent agreement of 89% and kappa statistic of 0.70. One hundred and sixty-eight studies were excluded as they did not meet all inclusion criteria. Reasons for exclusion included no prioritization using BWS ($n = 76$), no full text available (e.g., abstract only; $n = 38$), not published in peer-reviewed literature (e.g., dissertations; $n = 18$), secondary data analysis ($n = 15$), not having a defined population or using simulated data ($n = 19$), and full text not available in English ($n = 2$). Three hundred and eighty-two studies were included in the data extraction phase. After extraction, 217 studies were classified as in non-health

domains and excluded from this analysis. One hundred and sixty-five studies were included in the final analysis.

3.2 Application Type

Of the 165 studies meeting the eligibility criteria (Fig. 1), 34 studies were published prior to 2017 and 131 studies were published from 2017 through 2021. Several papers were ‘near-misses.’ These included papers reporting on the use of qualitative methods to evaluate the feasibility of BWS studies in patient populations with high medical need [50, 51]. Near-misses were identified in the full-text review stage.

The largest number of studies were published in 2021 ($n = 37$). Studies were published in a total of 97 journals, with 23 journals having published two or more studies included in our review. The journals with the highest number of BWS studies were: *PLoS One* ($n = 5$), *Patient Preference and Adherence* ($n = 4$), *The Patient-Patient Centered Outcomes Research* ($n = 4$), and *Journal of Medical Economics* ($n = 4$). Studies included in our review were conducted in six regions including North America, South America, Europe, Africa, Asia, and Oceania (Table 1). Terms study authors used to describe prioritization using BWS changed over time ($p = 0.04$). Most consistently “BWS” was used with no qualifier (67.6% vs 67.9%). Use of “MaxDiff” decreased from 20.6% to 9.2%, while use of BWS with the additional specification “object case” increased from 0% to 13.0%. Other terms used included BWS with the additional specification “case 1,” “best worst choice/task,” and “object scaling.” Across time periods, the overwhelming majority of studies (97.1% and 97.7%) reported an objective to provide substantive or empirical evidence on priorities ($p = 0.83$). Since 2017, fewer studies had educational or methodological objectives ($p = 0.03$), while there was a similar frequency of pilot studies ($p = 0.58$). For studies aimed to describe a variety of stakeholder perspectives; the most common perspective across both time periods was the patient/consumer (58.8% and 66.7%), followed by the provider/producer (38.2% and 32.8%), and the citizen/societal (5.9% and 7.6%). There were no significant differences any perspective across time periods (Fig. 2).

3.3 Development

The largest change in methods used for the development of BWS instruments between the two time periods (Table 2) is the increased use of pretesting (8.8% vs 24.4%; $p = 0.05$). There has been an increase in the use of literature reviews (52.9% vs 62.6%; $p = 0.30$), and key informant interviews (20.6% vs 35.6%; $p = 0.10$), although not statistically significant. Other common methods for instrument development across time periods included pilot testing, formal qualitative research, prior preference research, and existing lists of

objects. These were not mutually exclusive categories and 5% of studies reported the use of two or more methods for instrument development.

3.4 Design

Modes of survey administration (Table 3) included online, self-administration (paper returned via mail or in-person), and administration by a member of the study team. Since 2017, there were significant increases in the use of online administration (46.9% vs 76.9%; $p < 0.01$) and decreases in the use of traditional self-administration (31.2% vs 13.8%; $p = 0.02$). The time horizon used to frame the context of the BWS choice was most commonly the present, and its use increased since 2017 (76.5% vs 92.4%) while the use of future choices (14.7% vs 6.9%) and past choices (8.8% vs 0.8%) decreased ($p < 0.01$).

Across both time periods, the most common BWS experimental designs used were BIBDs with 52.9% before 2017 and 48.9% of studies since 2017 using BIBDs. Although there was no statistically significant difference between the two time periods ($p = 0.18$), there was an increase in designs generated by Sawtooth (17.6% vs 23.7%) and decreases in orthogonal designs (8.8% vs 4.6%) and other designs (8.8% vs 1.5%). Anchors for choice tasks changed over time ($p = 0.04$) with most/least” increasing in use (76.5% vs 89.3%), while “best/worst” (20.6% vs 10.7%) and other anchors descriptions decreased (2.9% vs 0.0%). Scales used to measure outcomes did not change significantly over time ($p = 0.48$). Importance/priorities remained the most common scale (79.4% vs 69.5%), despite its decrease in use and despite the increased use of preference (11.8% vs 14.5%) and emotion (8.8% vs 16.0%).

The total number of objects included in designs ranged from 5 to 60. Studies most commonly included 11 objects, followed by 16 objects and ten objects (Fig. 3). There was no significant difference in the average number of total objects in designs between the two time periods of interest (15.3 vs 16.0; $p = 0.68$). The number of objects presented per BWS task ranged from 3 to 11, and the most common number of objects per task was 4 and 5 (Fig. 4). There was no significant difference in the average number of objects presented per BWS task between the two time periods of interest (4.8 vs 4.6; $p = 0.40$). The mean total number of tasks also did not differ significantly between time periods (20.8 vs 24.3; $p = 0.60$). The mean number of BWS choice tasks presented to participants was generally consistent across time periods (13.1 vs 12.5; $p = 0.63$), and ranged from 1 to 50 across all studies, with half of all studies having between 10 and 16 choice tasks per respondent.

3.5 Administration/Analysis

Sample sizes ranged from as low as 15 to as high as 9289. The average sample size increased from 221.8 to 472.4 ($p = 0.11$) (Table 4). Three studies conducted between 2017 and 2020 had very large sample sizes of 2403, 3039, and 9289. In both time periods, the majority of studies provided no justification for their choice of sample size (76.5% and 65.6%). Though not significant ($p = 0.68$), justifications for sample size increased over time, including historical or empirical justification (11.8% vs 16.8%), rule of thumb (8.8% vs 12.2%), and formal sample size calculations (2.9% vs 5.3%).

Sawtooth was the most commonly used statistical program for data analysis in both time periods (35.3% and 27.5%). Since 2017, the use of other programs has increased including Stata (11.8% vs 22.1%), R (2.9% vs 15.3%), NLOGIT (0% vs 7.6%), Latent Gold Choice (0% vs 3.1%), and Excel (0% vs 6.1%), though none of these changes was significant. In both time periods, the most common approaches for analysis were probability/ratio rescaling (50.0% and 55.7%), regression coefficients (26.5% and 42.0%), and “B-W scores” (44.1% and 34.4%). Since 2017, significantly more studies conducted heterogeneity analyses (55.9% vs 75.6%; $p = 0.02$). The most common approaches for a heterogeneity analysis across time periods were stratification (47.1% and 53.4%), segmentation/latent class analysis (8.8% and 13.7%), and mixed logit (2.9% and 10.7%).

3.6 Evaluation of the PREFS Measure

Averages for total PREFS scores did not significantly differ between the two time periods (3.1 vs 3.1; $p = 0.98$), but average scores for two individual PREFS items did differ between the two periods (Table 5). Scores for whether the purpose was clearly stated increased significantly (91.2% vs 99.2%; $p < 0.01$), while scores for whether respondents are similar to nonrespondents decreased significantly (47.1% vs 24.4%; $p = 0.01$). No differences were observed in the explanation, findings, and significance items of the PREFS instrument across the two time periods. Quality scores measured via the subjective quality score improved from an average of 6.0 to an average of 6.7 ($p = 0.06$). Policy relevance scores were consistent between the two time periods (6.3 vs 6.9; $p = 0.13$).

Subjective quality score was positively associated with PREFS (coefficient = 1.305, $p < 0.001$; Table 6a). Among the individual PREFS items (Table 6b), using significance tests (1.859, $p < 0.001$) and having a clear explanation of methods (1.679, $p < 0.001$) had the strongest association with the subjective quality score, followed by findings (1.074, $p < 0.001$), respondents (0.721, $p = 0.001$), and purpose (0.682, $p = 0.588$). Controlling for the association of PREFS (Table 6c), other variables were also associated

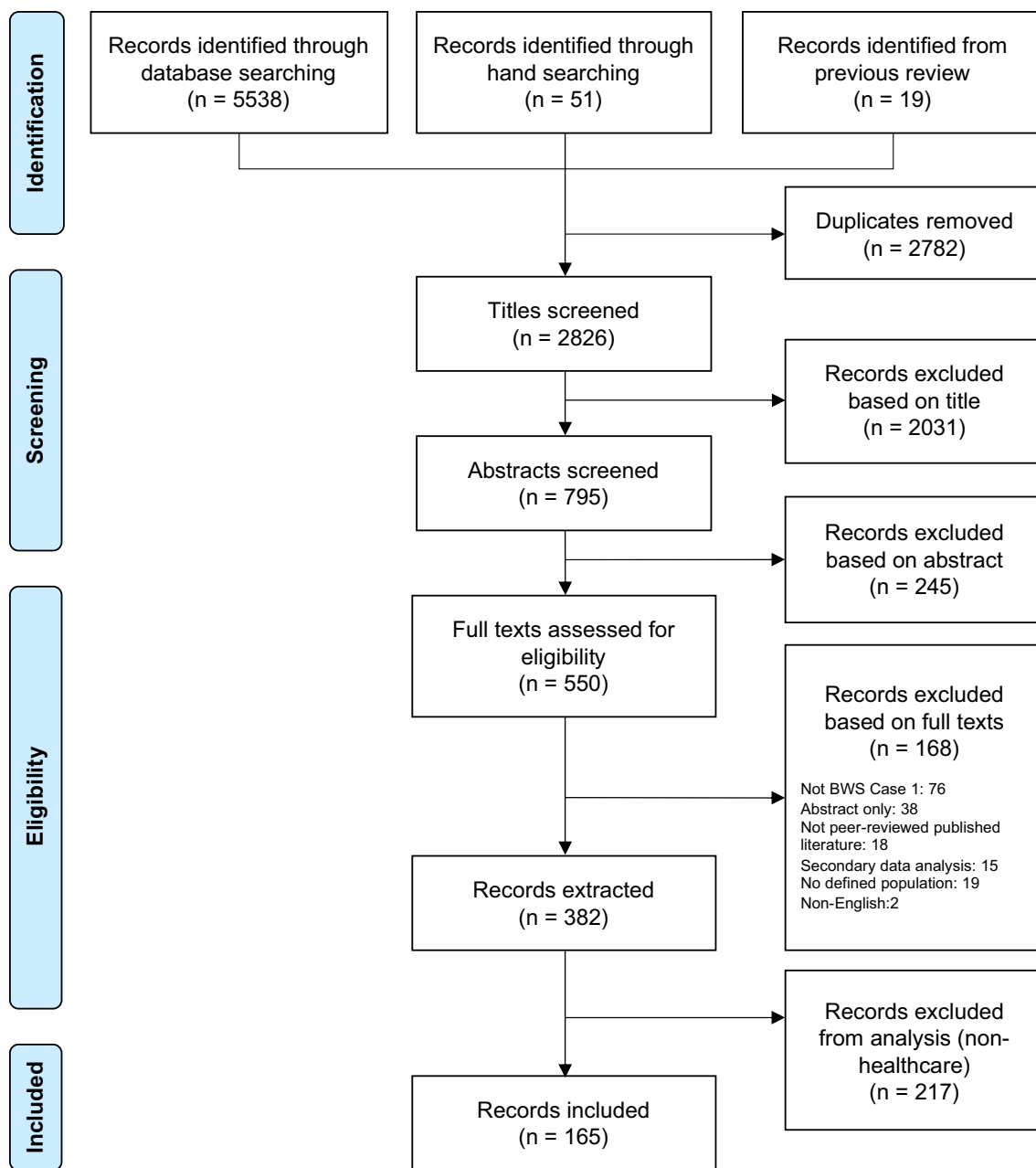
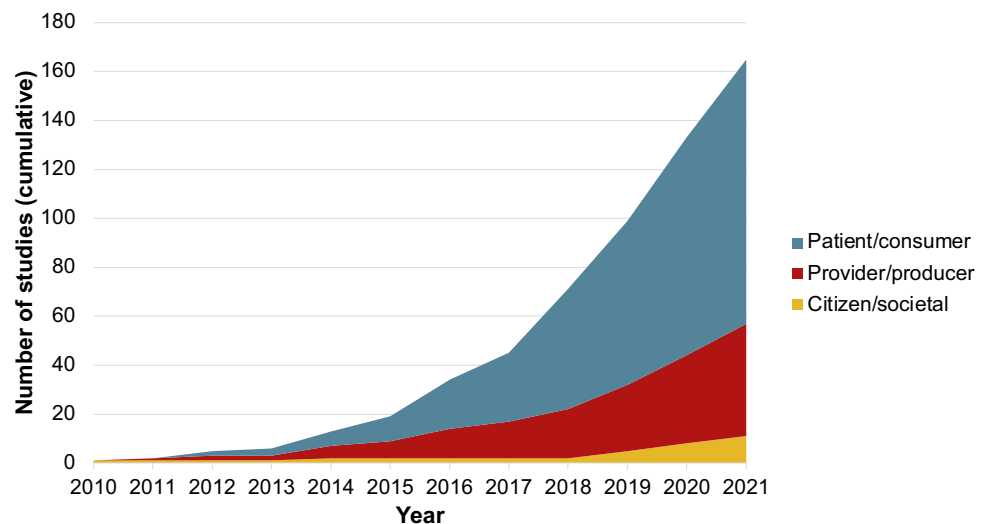


Fig. 1 Flow chart. *BWS* best-worst scaling

with subjective quality including policy relevance (0.174, $p = 0.001$), using a BIBD (0.741, $p < 0.001$), not reporting any development methods (-1.950 , $p = 0.020$), and sample size (-0.023 , $p = 0.0001$). On average, studies that used a heterogeneity analysis had higher subjective quality scores, but this effect was not significant (0.458, $p = 0.052$). Similar associations were observed when controlling for individual PREFS items (Table 6d), and this model had the highest explanatory power ($R^2 = 0.57$). The only change observed in the final model was that not reporting development methods was no longer significant (-0.851 , $p = 0.082$).

Policy relevance was associated with sample size (0.042, $p < 0.001$), after adjusting for PREFS, but not with using heterogeneity analysis, not reporting any developmental methods, or using a BIBD (Table 7a). Similar associations were observed when adjusting for subjective quality (Table 7b). Both models had low explanatory power ($R^2 = 0.06$ and 0.13 , respectively). The PREFS score was not associated with policy relevance after adjusting for subjective quality (Table 7c). PREFS was not associated with using heterogeneity analysis, not using developmental methods, using a BIBD, or sample size (Table 7d). Adding these

Fig. 2 Perspective of best–worst scaling studies over time**Table 2** Changes in the development of best–worst scaling instruments in health over time

	2010–16 <i>n</i> = 34	2017–21 <i>n</i> = 131	<i>P</i> value
Methods of instrument development*, <i>n</i> (%)			
Literature review	18 (52.9%)	82 (62.6%)	0.30
Pilot test	13 (38.2%)	37 (28.2%)	0.26
Pretest	3 (8.8%)	32 (24.4%)	0.05
Formal qualitative research	15 (44.1%)	51 (38.6%)	0.56
Key informant interviews	7 (20.6%)	47 (35.6%)	0.10
Prior preference research	3 (8.8%)	15 (11.5%)	0.66
Existing list of objects	2 (5.9%)	8 (6.1%)	0.96

*All that apply

variables to the model (model d) also did not greatly increase its explanatory power relative to the model only exploring policy relevance and subjective quality (model c) [$R^2 = 0.44$ and 0.42 , respectively].

4 Discussion

In this systematic review, we characterize the uses of BWS for measuring priorities in health, identify trends in the application of BWS in health settings over time, and assess the quality and policy relevance of these applications. We find that the popularity of BWS to measure priorities is growing, and more than three times as many applications of this method were published from 2017–21 ($n = 131$) compared with the period prior to 2017 ($n = 34$). We also explore what factors of BWS studies contribute to their quality and relevance to decision makers.

The use of consistent terminology for describing a methodological approach is an important factor in improving its acceptability, understanding, and opportunity for practical applications. Within the health literature, five different terms were used to describe prioritization using BWS. More than three quarters of studies use the term ‘BWS’ to refer to this method. We endorse the use of this terminology because it best reflects the method’s intention without incorporating analytic assumptions. MaxDiff is a potential misnomer in many cases because it implies an assumption about how respondents choose; it assumes respondents evaluate all pairwise comparisons and choose the best and worst based on the pair that reflects the maximum difference. The use of consistent terminology for describing BWS for prioritization and for BWS more generally (all types) is a worthwhile goal, albeit outside the scope of this paper.

We observed an increase in the use of literature reviews and key informant interviews to inform the development of BWS instruments over the two time periods. We found that only six studies (< 5%) did not report instrument development methods at all [52–57], which indicates a lack of reporting is not as problematic as we expected. However, under-reporting instrument development methods may still be problematic owing to the lack of detail reported about instrument development [45,58–62]. As a result, many authors question how to develop instruments (i.e., using literature reviews or formal qualitative methods) and how to report on their development work. Guidelines for transparent reporting of development work have been published [63], albeit in 2020 near the end of the current study’s period. Future reviews will determine whether instrument development methodology is more uniformly reported.

Balanced incomplete block design was the most frequently used experimental design for BWS. Balanced incomplete block design had half as many tasks on average as

Table 3 Changes in the design of best-worst scaling applications in health over time

	2010–2016 <i>n</i> = 34	2017–2021 <i>n</i> = 131	<i>P</i> value
Mode of survey administration*, <i>n</i> (%)			
Online	15 (46.9%)	100 (76.9%)	< 0.01
Self-administered** (paper)	10 (31.2%)	18 (13.8%)	0.02
Administered	7 (21.9%)	17 (13.0%)	0.20
Time-horizon, <i>n</i> (%)			
Present	26 (76.5%)	121 (92.4%)	< 0.01
Future	5 (14.7%)	9 (6.9%)	
Past	3 (8.8%)	1 (0.8%)	
Measurement scale, <i>n</i> (%)			
Importance/priorities	27 (79.4%)	91 (69.5%)	0.48
Preference	4 (11.8%)	19 (14.5%)	
Emotion	3 (8.8%)	21 (16.0%)	
Experimental design, <i>n</i> (%)			
BIBD	18 (52.9%)	64 (48.9%)	0.18
Sawtooth	6 (17.6%)	31 (23.7%)	
Orthogonal	3 (8.8%)	6 (4.6%)	
Random	0 (0.0%)	2 (1.5%)	
Other	3 (8.8%)	2 (1.5%)	
Not specified	4 (11.8%)	26 (19.5%)	
BWS anchor description, <i>n</i> (%)			
Most/least	26 (76.5%)	117 (89.3%)	0.04
Best/worst	7 (20.6%)	14 (10.7%)	
Other	1 (2.9%)	0 (0.0%)	
Objects total, mean (SD)	15.3 (9.0)	16.0 (8.9)	0.68
Objects per task, mean (SD)	4.8 (1.6)	4.6 (1.1)	0.40
Choice tasks total, mean (SD)	20.8 (23.9)	24.3 (34.7)	0.60
Choice tasks per respondent, mean (SD)	13.1 (3.3)	12.5 (6.4)	0.63

BIBD balanced incomplete block design, *BWS* best–worst scaling, *SD* standard deviation

*All that apply

**Self-administered a paper survey in person or at home and returned via mail

Fig. 3 Number of objects included in best–worst scaling studies



compared with non-BIBD studies (15.8 vs 33.2, $p < 0.01$) and tended to have fewer objects than non-BIBD studies (12.9 vs 18.9, $p < 0.01$). Balanced incomplete block designs are now widely available through Stata programs. We also observed a movement toward the use of Sawtooth

experiment designs since 2017. This may be a result of the ease with which they can be used. It is a positive finding that the percentage of studies reporting the use of orthogonal designs has decreased since 2017. This decrease is expected as they are now seen as inappropriate for use in BWS given

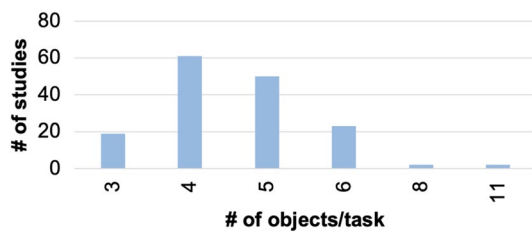


Fig. 4 Number of objects per task included in best–worst scaling studies

their non-uniform set sizes, and as access to BIBDs has increased. The use of orthogonal designs in BWS is likely a carry-over from their use in 2^K experiments, which were used prior to the introduction of BWS. A concerning finding is that 20% of studies did not report the experimental design used. Studies that did not specify an experimental design had a lower subjective quality score (5.2 vs 6.8, $p < 0.01$) than those that did. No differences in PREFS or policy relevance were identified between those reporting and not reporting on their experimental design. It is critical that study authors specify the experimental design and other essential design details in their methods for transparency, reproducibility, and credibility of their findings.

Sample sizes in BWS studies are increasing over time. While online administration and the use of large online panels may increase sample sizes, it comes at the risk of generating lower quality data. With the majority of studies not justifying their sample size choice, it is even more difficult to ascertain if that risk of low-quality data is warranted.

It is interesting to see how our findings compared to the most recent systematic review of DCEs in health [58]. While our data are not directly comparable (because of differences in timeframes and the relative scope of DCEs and BWS used for prioritization), some key differences can be observed (Table 8). Discrete choice experiments are more common than using BWS for prioritization (60 vs 26 per year). As expected, BWS can be used to compare a greater number of topics at one time; the average DCE has five attributes, while the BWS are used to prioritize on average 12 objects (and 60 objects). This said, the median number of choice tasks was the same at 12. Other key differences between the methods can be observed (Table 8), but more research is needed to directly compare different methods to measure both priorities and preferences.

We introduced and evaluated a global measure to assess subjective quality. Global measures are easy to use and can reduce evaluator burden as compared to lengthier, multi-item critical appraisal tools [64]. However, subjectivity can be problematic for quality evaluation in fields that lack consensus on what constitutes quality.

We acknowledge the limitations of the PREFS scale as a measure of study quality [41, 65, 66]. However, findings

from the current study paint a more promising picture of PREFS' validity than have other systematic reviews [67]. Furthermore, PREFS has some utility as it allows comparisons to previously published norms for other methods and health-focused literature. In the current research, we observed a significant positive association between PREFS and subjective quality. The observed association between PREFS and subjective quality may be biased because of the proximity of their report on the data extraction tool in the present study, as reviewers first evaluated PREFS and immediately after were asked to provide an overall quality score. The only individual PREFS item that was not associated with subjective quality was purpose, which is logically consistent with the fact that over 97% of studies report purpose and therefore not a source of discriminating quality.

The significant increase in the proportion of studies addressing the heterogeneity of preferences is a positive indicator of the quality of studies. Surprisingly, addressing heterogeneity was not significantly associated with the subjective quality score. The fact that three-quarters of studies in the more recent period included a heterogeneity analysis is a promising indicator for policy relevance as well. For instance, decision makers such as the US Food and Drug Administration have acknowledged that for preference measurement to be useful in regulatory decision making, it is important to understand how preferences may vary across observable and unobservable factors [68]. Heterogeneity analysis was not significantly associated with the policy relevance measure. A potential reason for this may be because reviewers focused on the policy relevance of study objectives and the context in which priorities could play a role to inform decision making and were not factoring in how the study results are presented into policy relevance.

There is great potential for BWS studies to influence policy given their focus on prioritization. Policymakers must weigh competing policy options and prioritize certain issues within healthcare. This often requires decisions to be made among people with competing goals. Best–worst scaling may serve as a prerequisite to formal frameworks for priority setting because of its ability to describe priorities and provide relevant information to decision makers [69–72]. A lack of information on priorities may impede the use of formal decision-making processes, which is intermittent at best [73]. Clearly articulated priorities may allow organizations to move away from non-explicit and informal decision making and move towards more legitimate, equitable, and efficient decision making.

In separate models, both PREFS and subjective quality were associated with policy relevance. This may indicate an overlap in the factors reviewers considered in assigning a score. In other words, quality may have been a consideration in policy relevance. However, in models in which PREFS was the outcome variable, policy relevance scores

Table 4 Changes in the administration/analysis of best–worst scaling applications studies in health over time

	2010–2016 <i>n</i> = 34	2017–2021 <i>n</i> = 131	<i>P</i> value
Sample size, mean (SD)	221.8 (205.6)	472.4 (896.2)	0.11
Sample size justification, <i>n</i> (%)			0.68
Rule of thumb	3 (8.8%)	16 (12.2%)	
Historical/empirical justification	4 (11.8%)	22 (16.8%)	
Sample size calculation	1 (2.9%)	7 (5.3%)	
Not specified	26 (76.5%)	86 (65.6%)	
Analytic program*, <i>n</i> (%)			
Sawtooth	12 (35.3%)	36 (27.5%)	0.37
Stata	4 (11.8%)	29 (22.1%)	0.18
<i>R</i>	1 (2.9%)	20 (15.3%)	0.06
NLOGIT	0 (0.0%)	10 (7.6%)	0.10
SPSS	4 (11.8%)	15 (11.5%)	0.96
Latent Gold Choice	0 (0.0%)	4 (3.1%)	0.30
SAS	2 (5.9%)	12 (9.2%)	0.54
Excel	0 (0.0%)	8 (6.1%)	0.14
Analytic approach*, <i>n</i> (%)			
Probability/ratio rescaling	17 (50.0%)	73 (55.7%)	0.55
B–W score	15 (44.1%)	45 (34.4%)	0.29
Coefficients	9 (26.5%)	55 (42.0%)	0.10
Counts	8 (23.5%)	27 (20.6%)	0.71
Square roots	5 (14.7%)	8 (6.1%)	0.10
SUCRA	0 (0.0%)	2 (1.5%)	0.47
Theoretical assumption, <i>n</i> (%)			0.42
MaxDiff	12 (35.3%)	33 (25.2%)	
Sequential BWS	3 (8.8%)	18 (13.7%)	
Simultaneous	1 (2.9%)	1 (0.8%)	
Not specified	18 (52.9%)	79 (60.3%)	
Directionality (unidirectional), <i>n</i> (%)	21 (61.8%)	73 (55.7%)	0.53
Heterogeneity analysis used, <i>n</i> (%)	19 (55.9%)	99 (75.6%)	0.02
Heterogeneity analysis method*, <i>n</i> (%)			
Stratification	16 (47.1%)	70 (53.4%)	0.51
Segmentation/latent class	3 (8.8%)	18 (13.7%)	0.44
Mixed logit	1 (2.9%)	14 (10.7%)	0.16
Individual level score	1 (2.9%)	3 (2.3%)	0.82

SD standard deviation, *SUCRA* Surface Under the Cumulative Ranking score

*All that apply

were not statistically significant, which indicates that policy relevance may not have been a factor considered in assigning PREFS scores. This result is consistent with the fact that the individual items that are included in PREFS, and therefore the items that prime reviewers before assigning scores, do not include policy relevance.

We acknowledge that our policy relevance scores have conceptual and empirical issues. We have not tried to validate this subjective measure and have concerns about the validity and reliability of such a metric in the absence of an underlying theory. This said, increasing the policy relevance of prioritization studies in health is a major contemporary issue and our measure is a first of its type. Further, policy

relevance scores were intended to measure the potential relevance of a study's research question to policy or decision making. However, we were unable to ascertain the extent to which priority information obtained from BWS studies is actually used in decision making.

This systematic review has several strengths. A robust search strategy was used covering seven databases and hand searching. A series of quality checks were implemented to ensure data quality including a group review for consensus, cross-validation by one author, and pilot testing of the extraction instrument. There are limitations inherent to the systematic review process. Our data are limited to what is published in manuscripts and therefore subject to publication

Table 5 Changes in the quality and policy relevance of best–worst scaling studies in health over time

	2010–2016 <i>n</i> = 34	2017–2021 <i>n</i> = 131	<i>P</i> value
PREFS total score, mean (SD)	3.1 (0.9)	3.1 (0.9)	0.98
Purpose clearly stated, <i>n</i> (%)	31 (91.2%)	130 (99.2%)	<0.01
Respondents similar to non-respondents, <i>n</i> (%)	16 (47.1%)	32 (24.4%)	0.01
Explanation of methods clear, <i>n</i> (%)	21 (61.8%)	92 (70.2%)	0.34
Findings include all respondents, <i>n</i> (%)	9 (26.5%)	41 (31.3%)	0.59
Significance tests were used, <i>n</i> (%)	27 (79.4%)	105 (80.2%)	0.92
Subjective quality score, mean (SD)	6.0 (1.7)	6.7 (1.8)	0.06
Policy relevance score, mean (SD)	6.3 (2.0)	6.9 (1.9)	0.13

SD standard deviation

Table 6 Determinants of subjective quality in best–worst scaling studies (*n* = 165)

Variable	Subjective quality (a)	Subjective quality (b)	Subjective quality (c)	Subjective quality (d)
PREFS	1.305* (0.12)		1.113* (0.12)	
Purpose of study		0.682 (1.25)		0.490 (0.81)
Respondent sampling		0.721* (0.22)		0.573* (0.21)
Explanation of assessment methods		1.679* (0.23)		1.480* (0.21)
Findings (complete sets of responses included)		1.074* (0.22)		0.934* (0.21)
Significance test use		1.859* (0.30)		1.629* (0.27)
Policy relevance			0.174* (0.05)	0.192* (0.05)
Heterogeneity analysis			0.458 (0.23)	0.352 (0.22)
No development methods			−1.195* (0.50)	−0.851 (0.49)
BIBD			0.741* (0.20)	0.748* (0.20)
Sample size (100)			−0.023* (0.01)	−0.021* (0.01)
Constant	2.546* (0.40)	2.703* (1.26)	1.401* (0.40)	1.488 (0.86)
<i>R</i> ²	0.42	0.47	0.53	0.57

BIBD balanced incomplete block design, **p* < 0.05

Ordinary least squares performed on the full data set of all studies in both time periods

Robust standard errors in parentheses

Table 7 Determinants of policy relevance and PREFS in best–worst scaling studies (*n* = 165)

Variable	Policy relevance (a)	Policy relevance (b)	PREFS (c)	PREFS (d)
PREFS	0.404* (0.17)			
Subjective quality		0.362* (0.09)	0.323* (0.02)	0.329* (0.03)
Policy relevance			0.000 (0.03)	−0.005 (0.03)
Heterogeneity analysis	0.116 (0.37)	−0.069 (0.34)		0.142 (0.13)
No development methods	0.212 (0.63)	0.645 (0.65)		0.024 (0.13)
BIBD	−0.133 (0.30)	−0.394 (0.30)		−0.208 (0.11)
Sample size (100)	0.042* (0.01)	0.047* (0.01)		−0.001 (0.00)
Constant	5.313* (0.62)	4.404* (0.63)	0.943* (0.22)	0.942* (0.23)
<i>R</i> ²	0.06	0.13	0.42	0.44

BIBD balanced incomplete block design, **p* < 0.05

Ordinary least squares performed on the full data set of all studies in both time periods

Robust standard errors in parentheses

Table 8 Comparison of best-worst scalings and discrete-choice experiments

	DCE 2013–2017 <i>n</i> = 301	BWS 2017–2021 <i>n</i> = 131
Number of studies per year, mean (range)	60 (32–98)	26 (11–37)
Used of formal qualitative methods, <i>n</i> (%)	258 (86%)	51 (39%)
Number of attributes (DCE) or objects (BWS), median (range)	5 (2–21)	12 (5–60)
Number of choice tasks per respondent, median (range)	12 (1–32)	12 (1–300)
Sample size, median (range)	401 (35–30,6000)	220 (15–9289)
Country/location of study, <i>n</i> (%)		
USA	50 (17%)	67 (41%)
EU	72 (24%)	25 (19%)
UK	50 (17%)	5 (3%)
Australia	30 (10%)	5 (3%)
Canada	25 (8%)	4 (2%)
Other/not specified	102 (34%)	25 (32%)
Mode of survey administration*, <i>n</i> (%)		
Online	172 (57%)	100 (77%)
Self-administered (paper)	69 (23%)	18 (14%)
Administered	44 (15%)	17 (13%)
Other/not specified	16 (5%)	0 (0%)
Perspective*, <i>n</i> (%)		
Patient/consumer	110 (37%)	88 (67%)
Provider/producer	39 (13%)	43 (33%)
Citizen/societal	81 (27%)	10 (8%)
Other/not specified	98 (33%)	0 (0%)
Analytic program*, <i>n</i> (%)		
Stata	94 (31%)	29 (22%)
Sawtooth	16 (5%)	36 (28%)
NLOGIT	65 (22%)	10 (8%)
<i>R</i>	10 (3%)	20 (15%)
SAS	17 (6%)	12 (9%)
Other/not specified	99 (33%)	0 (0%)

BWS best–worst scaling, *DCE* discrete-choice experiment, *EU* European Union

*All that apply

bias. Systematic reviews aim to synthesize a body of literature and evaluate the quality of the studies in that literature. It is difficult to synthesize, qualitatively or quantitatively, studies that are as varied in their outcomes as what we found. It is also difficult to evaluate quality without a consensus definition of quality or a good evaluative instrument.

Our paper has not clearly determined the best practices for the application of BWS in health. This said, our efforts to document norms and standards for the method can facilitate a great awareness of this method and can help with placing an individual contribution in the context of the larger literature. Furthermore, given the number of different methods available to study preferences and priorities, it may not be feasible to have the degree of a consensus-based standard for every method. Several contemporary efforts are focused

on developing standards for a wide array of preference and prioritization methods [30, 74].

Our paper is focused only on health, and as such does not focus on the application of BWS in other literature. As is implicit in our search strategy, we have ongoing research to explore the application of BWS in fields other than health and compare health and non-health studies using BWS.

5 Conclusions

This study provides the most detailed review of an emerging class of methods that aim to quantify the priorities of patients and other stakeholders in health. Best–worst scaling provides researchers with a theory-driven method to evaluate priorities that allows for both ordinal and cardinal valuations

of issues that matter most. We provide evidence that the application of BWS has been rapidly expanding over the last decade and detail how best research practices have been evolving over time. Our study provides important evidence for researchers, reviewers, and readers of BWS studies on how this method is being used in health. We also have made contributions to the assessment of the quality and policy relevance of these studies that are relevant to those who want to apply BWS and to those interested in the broader assessment of a range of methods used to study the priorities and preferences of patients and other stakeholders in health.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40273-022-01167-1>.

Acknowledgements The authors acknowledge Sarah M. Hyman for her research assistance in developing and testing the literature search strategy and building the database.

Declarations

Funding John F.P. Bridges holds an Innovation in Regulatory Science Award from the Burroughs Wellcome Fund. Ilene L. Hollin, Jonathan Paskett, and Norah Crossnohere are supported, in part, by this award.

Conflict of interest/competing interests Ilene L. Hollin, Jonathan Paskett, Anne L.R. Schuster, Norah L. Crossnohere, and John F.P. Bridges have no conflicts of interest that are directly relevant to the content of this article.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material Not applicable.

Code availability Not applicable.

Author contributions Conceptualization: JFPB; methodology: ILH, JFPB, NLC; literature search: JP; data extraction: ILH, JP, ALRS, NLC, JFPB; data analysis: JFPB, JP, NLC; writing, original draft preparation: ILH, JFPB; writing, critical review and editing: ILH, JP, ALRS, NLC, JFPB; funding acquisition: JFPB; supervision: JFPB.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc/4.0/>.

References

1. Bridges JF, Selck FW, Gray GE, McIntyre JA, Martinson NA. Condom avoidance and determinants of demand for male circumcision in Johannesburg, South Africa. *Health Policy Plan.* 2011;26(4):298–306.
2. Bridges JF, Gallego G, Kudo M, et al. Identifying and prioritizing strategies for comprehensive liver cancer control in Asia. *BMC Health Serv Res.* 2011;11(1):1–12.
3. Peay H, Hollin I, Bridges J. Prioritizing parental worry associated with Duchenne muscular dystrophy using best–worst scaling. *J Genet Counsel.* 2016;25(2):305–13.
4. Tsai JH, Scheimann AO, McCandless SE, Strong TV, Bridges JFP. Caregiver priorities for endpoints to evaluate treatments for Prader-Willi syndrome: a best–worst scaling. *J Med Econ.* 2018;21(12):1230–7.
5. Oakes AH, Garmo VS, Bone LR, Longo DR, Segal JB, Bridges JFP. Identifying and prioritizing the barriers and facilitators to the self-management of type 2 diabetes mellitus: a community-centered approach. *Patient.* 2017;10(6):773–83.
6. Ejaz A, Spolverato G, Bridges JF, Amini N, Kim Y, Pawlik TM. Choosing a cancer surgeon: analyzing factors in patient decision making using a best–worst scaling methodology. *Ann Surg Oncol.* 2014;21(12):3732–8.
7. Gallego G, Bridges JF, Flynn T, Blauvelt BM, Niessen LW. Using best–worst scaling in horizon scanning for hepatocellular carcinoma technologies. *Int J Technol Assessm Health Care.* 2012;28(3):339–46.
8. Witteman HO, Ndjaboue R, Vaisson G, et al. Clarifying values: an updated and expanded systematic review and meta-analysis. *Med Decis Making.* 2021;41(7):801–20.
9. Chalasani M, Vaidya P, Mullin T. Enhancing the incorporation of the patient's voice in drug development and evaluation. *Res Invol Engagem.* 2018;4:10.
10. Forsythe LP, Ellis LE, Edmundson L, et al. Patient and stakeholder engagement in the PCORI pilot projects: description and lessons learned. *J Gen Intern Med.* 2016;31(1):13–21.
11. Bridges J. Stated preference methods in health care evaluation: an emerging methodological paradigm in health economics. *Appl Health Econ Health Policy.* 2003;2(4):213–24.
12. Committee on Quality of Health Care in America. Institute of Medicine, editor. *Crossing the quality chasm: a new health system for the 21st century.* Washington, DC: National Academy Press; 2001.
13. Janus SI, Weernink MG, van Til JA, Raisch DW, van Manen JG, IJzerman MJ. A systematic review to identify the use of preference elicitation methods in health care decision making. *Value Health.* 2014;17(7):515–6.
14. Bridges JF, Slawik L, Schmeding A, Reimer J, Naber D, Kuhnigk O. A test of concordance between patient and psychiatrist valuations of multiple treatment goals for schizophrenia. *Health Expect.* 2013;16(2):164–76.
15. Ryan M, Farrar S. Using conjoint analysis to elicit preferences for health care. *BMJ.* 2000;320(7248):1530–3.
16. 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.
17. Mühlbacher AC, Kaczynski A, Zweifel P, Johnson FR. Experimental measurement of preferences in health and healthcare using best–worst scaling: an overview. *Health Econ Rev.* 2016;6(1):2.
18. Cheung KL, Wijnen BF, Hollin IL, et al. Using best–worst scaling to investigate preferences in health care. *Pharmacoeconomics.* 2016;34(12):1195–209.

19. Vass C, Gray E, Payne K. Discrete choice experiments of pharmacy services: a systematic review. *Int J Clin Pharm.* 2016;38(3):620–30.
20. Mandeville KL, Lagarde M, Hanson K. The use of discrete choice experiments to inform health workforce policy: a systematic review. *BMC Health Serv Res.* 2014;14(1):1–14.
21. Quaife M, Terris-Prestholt F, Di Tanna GL, Vickerman P. How well do discrete choice experiments predict health choices? A systematic review and meta-analysis of external validity. *Eur J Health Econ.* 2018;19(8):1053–66.
22. Marshall D, Bridges JF, Hauber B, et al. Conjoint analysis applications in health: how are studies being designed and reported? An update on current practice in the published literature between 2005 and 2008. *Patient.* 2010;3(4):249–56.
23. Webb EJ, Meads D, Eskyte I, et al. A systematic review of discrete-choice experiments and conjoint analysis studies in people with multiple sclerosis. *Patient.* 2018;11(4):391–402.
24. Cheung KL, Wijnen BF, Hollin IL, et al. Using best–worst scaling to investigate preferences in health care. *Pharmacoeconomics.* 2016;34:1195–209.
25. Finn A, Louviere JJ. Determining the appropriate response to evidence of public concern: the case of food safety. *J Public Policy Marketing.* 1992;11(2):12–25.
26. Thurstone LL. A law of comparative judgment. *Psychol Rev.* 1927;34(4):273.
27. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. *Frontiers in econometrics.* New York: Academic Press; 1974. p. 105–42.
28. Campbell BF, Sengupta S, Santos C, Lorig KR. Balanced incomplete block design: description, case study, and implications for practice. *Health Educ Q.* 1995;22(2):201–10.
29. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making. *Pharmacoeconomics.* 2008;26(8):661–77.
30. Bridges JF, Hauber AB, Marshall D, et al. Conjoint analysis applications in health: a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value Health.* 2011;14(4):403–13.
31. Thokala P, Devlin N, Marsh K, et al. Multiple criteria decision analysis for health care decision making: a introduction: report 1 of the ISPOR MCDA Emerging Good Practices Task Force. *Value Health.* 2016;19(1):1–13.
32. Webb EJ, Meads D, Lynch Y, et al. Attribute selection for a discrete choice experiment incorporating a best–worst scaling survey. *Value Health.* 2021;24(4):575–84.
33. Louviere JJ, Flynn TN, Marley AAJ. *Best–worst scaling: theory, methods and applications.* Cambridge: Cambridge University Press; 2015.
34. Mateen F, Oh J, Tergas A, Bhayani N, Kamdar B. Titles versus titles and abstracts for initial screening of articles for systematic reviews. *Clin Epidemiol.* 2013;5:89–95.
35. 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.
36. Joy SM, Little E, Maruthur NM, Purnell TS, Bridges JFJP. Patient preferences for the treatment of type 2 diabetes: a scoping review. *Pharmacoeconomics.* 2013;31(10):877–92.
37. Showalter TN, Mishra MV, Bridges JF. Factors that influence patient preferences for prostate cancer management options: a systematic review. *Patient Prefer Adher.* 2015;9:899.
38. Zhou M, Thayer WM, Bridges JF. Using latent class analysis to model preference heterogeneity in health: a systematic review. *Pharmacoeconomics.* 2018;36(2):175–87.
39. Kinsella N, Stattin P, Cahill D, et al. Factors influencing men's choice of and adherence to active surveillance for low-risk prostate cancer: a mixed-method systematic review. *Eur Urol.* 2018;74(3):261–80.
40. Sain N, Willems D, Charokopou M, Hilgsmann M. The importance of understanding patient and physician preferences for psoriasis treatment characteristics: a systematic review of discrete-choice experiments. *Curr Med Res Opin.* 2020;36(8):1257–75.
41. Beckham SW, Crossnohere NL, Gross M, Bridges JFP. Eliciting preferences for HIV prevention technologies: a systematic review. *Patient.* 2020;14(2):151–74.
42. Valderas JM, Ferrer M, Mendivil J, et al. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. *Value Health.* 2008;11(4):700–8.
43. Bridges J et al. A framework for studying patient preferences to inform decision making in healthcare: an ISPOR Good Practices Task Force report. In development.
44. Johnson FR, Lancsar E, Marshall D, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. *Value Health.* 2013;16(1):3–13.
45. Vass C, Rigby D, Payne K. The role of qualitative research methods in discrete choice experiments: a systematic review and survey of authors. *Med Decis Making.* 2017;37(3):298–313.
46. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. *Patient.* 2015;8(5):373–84.
47. Crossnohere N, Janssen E, Knight SJ, Hauber B, Bridges J. MSR36 Developing the Consolidated Health Outcomes/Interventions Choice-Modeling Evaluation Standards (CHOICES): A Content Analysis. *Value Health.* 2022;25(7):S524–S525.
48. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med.* 2005;37(5):360–3.
49. Belur J, Tompson L, Thornton A, Simon M. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. *Sociol Methods Res.* 2021;50(2):837–65.
50. Bailey C, Kinghorn P, Hewison A, et al. Hospice patients' participation in choice experiments to value supportive care outcomes. *BMJ Support Palliat Care.* 2019;9(4): e377.
51. Beusterien K, Chan E, Such P, et al. Development of a stated-preference instrument to prioritize treatment goals in recent onset schizophrenia. *Curr Med Res Opin.* 2017;33(12):2129–36.
52. Kurkjian TJ, Kenkel JM, Sykes JM, Duffy SC. Impact of the current economy on facial aesthetic surgery. *Aesthet Surg J.* 2011;31(7):770–4.
53. Cozmuta R, Merkel PA, Wahl E, Fraenkel L. Variability of the impact of adverse events on physicians' decision making. *BMC Med Inform Decis Making.* 2014;14(1):1–7.
54. Narurkar V, Shamban A, Sissins P, Stonehouse A, Gallagher C. Facial treatment preferences in aesthetically aware women. *Dermatol Surg.* 2015;41:S153–60.
55. Jagdeo J, Keaney T, Narurkar V, Kolodziejczyk J, Gallagher CJ. Facial treatment preferences among aesthetically oriented men. *Dermatol Surg.* 2016;42(10):1155–63.
56. Essers I, Hilgsmann M, Kiltz U, et al. Development of one general and six country-specific algorithms to assess societal health utilities based on ASAS HI. *RMD Open.* 2019;5(1): e000872.
57. Bennett WL, Aschmann HE, Puhon MA, et al. A benefit–harm analysis of adding basal insulin vs sulfonylurea to metformin to manage type II diabetes mellitus in people with multiple chronic conditions. *J Clin Epidemiol.* 2019;113:92–100.
58. Soekhai V, de Bekker-Grob EW, Ellis A, Vass C. Discrete choice experiments in health economics: past, present and future. *Pharmacoeconomics.* 2019;37:201–26.
59. Coast J, Horrocks S. Developing attributes and levels for discrete choice experiments using qualitative methods. *J Health Serv Res Policy.* 2007;12(1):25–30.

60. Coast J, Al-Janabi H, Sutton EJ, et al. Using qualitative methods for attribute development for discrete choice experiments: issues and recommendations. *Health Econ.* 2012;21(6):730–41.
61. 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.
62. Louviere JJ, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. *Health Econ Policy Law.* 2009;4(Pt 4):527–46.
63. Hollin IL, Craig BM, Coast J, et al. Reporting formative qualitative research to support the development of quantitative preference study protocols and corresponding survey instruments: guidelines for authors and reviewers. *Patient.* 2020;13(1):121–36.
64. Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. *J Man Manip Ther.* 2009;17(3):163–70.
65. Brown L, Lee T-H, De Allegri M, Rao K, Bridges JF. Applying stated-preference methods to improve health systems in sub-Saharan Africa: a systematic review. *Expert Rev Pharmacoecon Outcomes Res.* 2017;17(5):441–58.
66. Lack A, Hiligsmann M, Bloem P, Tünneßen M, Hutubessy R. Parent, provider and vaccinee preferences for HPV vaccination: a systematic review of discrete choice experiments. *Vaccine.* 2020;38(46):7226–38.
67. Beckham SW, Crossnohere NL, Gross M, Bridges JF. Eliciting preferences for HIV prevention technologies: a systematic review. *Patient.* 2021;14(2):151–74.
68. US FDA. Patient preference information: voluntary submission, review in premarket approval applications, humanitarian device exemption applications, and de novo requests, and inclusion in decision summaries and device labeling: guidance for industry, food and drug administration staff, and other stakeholders. US Food and Drug Administration, Rockville, MD; 2016.
69. Logeman C, Guha C, Howell M, et al. Developing consensus-based outcome domains for trials in children and adolescents with CKD: an international Delphi survey. *Am J Kidney Dis.* 2020;76(4):533–45.
70. Cho Y, Rangan G, Logeman C, et al. Core outcome domains for trials in autosomal dominant polycystic kidney disease: an international Delphi survey. *Am J Kidney Dis.* 2020;76(3):361–73.
71. Yeboyo HG, Aschmann HE, Yu T, Puhana MA. Should statin guidelines consider patient preferences? Eliciting preferences of benefit and harm outcomes of statins for primary prevention of cardiovascular disease in the sub-Saharan African and European contexts. *BMC Cardiovasc Disord.* 2018;18(1):1–11.
72. Schoenborn NL, Crossnohere NL, Bridges JF, Pollack CE, Pilla SJ, Boyd CM. Patient perceptions of diabetes guideline frameworks for individualizing glycemic targets. *JAMA Intern Med.* 2019;179(12):1642–9.
73. Seixas BV, Dionne F, Mitton C. Practices of decision making in priority setting and resource allocation: a scoping review and narrative synthesis of existing frameworks. *Health Econ Rev.* 2021;11(1):2.
74. Hauber AB, González JM, Groothuis-Oudshoorn CGM, et al. Statistical methods for the analysis of discrete choice experiments: a report of the ISPOR Conjoint Analysis Good Research Practices Task Force. *Value Health.* 2016;19(4):300–15.