Simply, the Best-Worst Scaling method is more reliable than ranking

Garston Liang, Mackenzie Glover, and Guy E. Hawkins

School of Psychological Sciences, University of Newcastle, Australia

Author Note

Please address correspondence to Garston Liang (*garston.liang@gmail.com*) School of Psychological Science, University of Newcastle, Callaghan, Newcastle, Australia, 2308.

ORCID: 0000-0002-9230-7258

Abstract

Two popular methods of preference elicitation are rankings and Best-Worst Scaling (BWS). Rankings, while simple and widely adopted, can be burdensome with larger item sets and fail to capture indifference between options that are neither loved nor hated. Best-Worst Scaling is a survey method that sidesteps the set size problem by capitalising on people's natural capacity to identify preferences at the extremes. Across three experiments, our primary finding is that elicited preferences for *ranking* and *BWS* methods align, and that BWS methods provide additional resolution to resolve the indifference between middling options where rankings can struggle as well as the relative importance of each option. Moreover, we show that BWS methods exhibit greater test-retest reliability compared to rankings, even over time frames as short as minutes. Taken together, our results privilege BWS as a reliable and readily accessible alternative to ranking methods for preference elicitation.

Keywords: ranking, best-worst scaling, preference elicitation, reliability

Introduction

If Christmas dinner conversation ever needed more spice, arguing over the best era of music might just do it. All of a sudden, an ordinarily quiet uncle rambles on about The Beatles, music of someone's good-old-days is clearly the best, and modern pop music is unquestionably the worst. This might not be the most agreeable conversation between generations (Van Dam, 2024) but, perhaps, it is a familiar one where the range of preferences is immediately clear and how one asks about these preferences matters. Beyond the dinner table, eliciting preferences is a foundational component to understanding a range of complex decisions including health care choices, consumer preferences, and political polling (Cooper & Hawkins, 2019; Ryan, 2004; Viney et al., 2002).

Preference elicitation has a rich history in the psychological and behavioural economics literature (Edwards, 1954; Von Neumann & Morgenstern, 2007). A common theoretical framework is to assume that individuals have stable internal representations of their preferences and variability arises from the measurement. Therefore, the choice of elicitation method is centrally important as it provides a window into people's underlying representations, albeit with varying degrees of opacity. Given the many ways to elicit preferences, a critical question is to what extent do common methods agree with one another?

This paper compares two popular methods of preference elicitation, namely *ranking* and *Best-worst Scaling*. Ranking tasks are ubiquitous and favoured for their relative simplicity (M. D. Lee et al., 2012; Montgomery et al., 2024). Individuals order a list of items according to some criteria (e.g., most preferred to least preferred music genre, or vice versa). For smaller items sets, e.g., ranking 3 genres, rankings can be easily obtained from relatively few comparisons between the items. However, rankings can be muddied by indifference between options and, particularly for larger sets, the number of comparisons drastically increases, e.g., a top 10 list demands 45 pairwise comparisons (Isaac & Schindler, 2014).

Best-Worst Scaling, henceforth BWS, is a survey method that sidesteps this set size problem by capitalising on people's natural capacity to identify extreme preferences (Louviere et al., 2013). Briefly, BWS methods involve individuals indicating their most favoured (i.e., best) and least favoured option (i.e., worst) over multiple sets of options, where each set is composed of different options. Despite its simplicity, elicited preferences from BWS are generally consistent and respondents benefit from the cognitive ease of selecting between options at the extremes from relatively small sets of options (Marley & Louviere, 2005).

In this paper, we directly compare *ranking* and *BWS* across three experiments that asked participants about preferences between 10 life values (e.g., rank the importance of benevolence, stimulation, etc., Schwartz, 1994). To preface the results, our primary finding is that elicited preferences for *ranking* and *BWS* methods are congruent though BWS methods provide the additional resolution to discern indifferences between options where rankings cannot. Additionally, we show that BWS methods exhibit greater test-retest reliability compared to rankings, even within time frames as short as minutes. Taken together, our results privilege BWS as a reliable and readily accessible alternative to ranking methods[1].

Preference elicitation

Best-Worst Scaling is a preference elicitation method that capitalises on the cognitive ease of identifying preferences at the extremes (Louviere et al., 2015). In this paper, we consider case 1 'object' BWS methods. Typically, individuals are presented with a set of items and asked to indicate their most preferred *and* least preferred item within a set (example in Figure 1). Each item is independent, like a genre of music, and each set contains at least 3 possible alternatives. Repeating these 'best' and 'worst' choices across carefully balanced

[1] For the reader interested in using BWS themselves, we refer to Louviere et al. (2013) for an applied demonstration of BWS and Louviere et al. (2015) for a comprehensive guide.

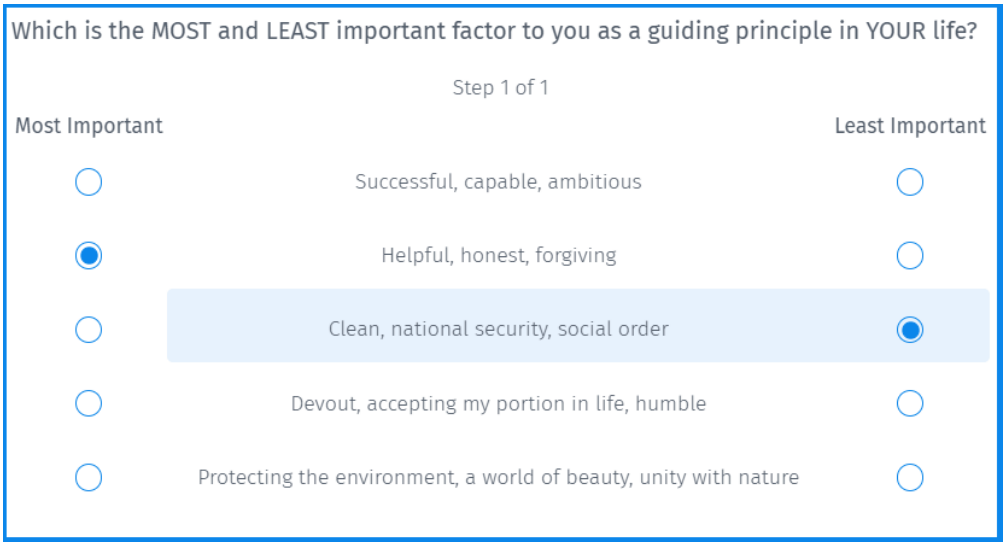

Figure 1. Example of a single Best-Worst Scaling choice set involving five life values with *item labels*, see Table 1 for full list. In the BWS experiments, participants completed 11 choice sets of best & worst choices, like the screenshot example, across changing subsets of life values. Across subsets, each life value is presented six times in total and each pair of values co-appeared three times. This value-presentation design was imported from the BWS adaptation of Schwartz's Values Theory used in J. A. Lee et al. (2008).

sets reveals for which items decision-makers have clear preferences and for which they entertain trade-offs.

Beyond ordinal preferences, a core advantage of BWS is that it compares apples with apples. Even when many important but qualitatively different factors are considered, each is placed along a common latent continuum and can be quantitatively weighed against one another. For example in health settings, Ejaz et al. (2014) asked oncology patients what information helped determine their choice of surgeon. While 'surgeon training' was centrally important to patient decisions, BWS methods provided the added resolution to identify that it was twice as important as 'surgeon experience' and over 3 times more important than the 'hospital's reputation'. Taken together, BWS provide rich data from relatively simple responses to understand the breadth of trade-offs in decision-maker

preferences.

In this domain, ranking tasks are a natural comparison. Ranking is quick, and, for respondents, the goal is immediately apparent even if the exact individual ranks are not. These advantages make ranking inherently appealing as a simple, ubiquitous, and relatively inexpensive manner to elicit ordinal preferences.

A common assumption in the preference elicitation literature is that any method provides imperfect access to the same underlying internal state of preferences. While the internal state remains consistent, expressed preferences might carry perturbations that can arise due to idiosyncratic response styles, such as when some people avoid the extremes of a rating scale (Grimmond et al., in press), or idiosyncrasies of the elicitation method itself, such as a ranking task that prevents expressions of equivalent preferences between options. This added randomness to expressed preferences means our comparison of ranking to BWS methods cannot inherently privilege either as closer to the 'true' state of individual's beliefs. Furthermore, like with music taste, there is often no 'ground truth' by which to compare subjective elicited preferences.

To remedy these complications, we incorporate three features into our experimental design. First, our investigations centered on preferences between life values. We chose this domain because values are personally relevant with natural diversity in preferences across people. Our experiments utilised a previously validated set of 10 life values based on a condensed BWS version of Schwartz' Value Theory (J. A. Lee et al., 2008; Schwartz, 1994). The main benefit for our investigation is that people's life values are typically stable and fundamental changes take place on the timescales of years dwarfing our experimental demands of mere minutes (Rokeach, 1973; Schwartz, 1992). We therefore expect strong test-retest reliability for both the ranking and BWS elicitation tools.

Second, our experiments investigate whether elicited preferences can be preserved when their descriptions are altered. Returning to our music analogy, consider that when describing the Beatles, one could define a broad genre (e.g., rock or pop) or their

characteristic qualities (e.g., iconic counter-melodies, chord progressions, vocal harmonies). The particular choice of description highlights distinct features across a common broader concept. Along a similar vein, our experiments describe the set of life values along broad *dimensions*, such as 'benevolence', as compared to example qualities henceforth called *items*, such as 'helpful, honest, & forgiving' (see Table 1). We anticipate that dimension- and item-descriptors will evoke different interpretations across individuals. Over and above these interpretations, however, we seek to understand whether rankings or BWS better preserves the elicited preferences across changes in description when both levels of description ostensibly target the same latent psychological concept.

Third, in all three experiments we elicit preferences from individuals twice. From a random utility perspective, preference stability over time is one indicator of validity in the absence of ground truth. To control for changes over time, we begin by experimentally determining the test-retest reliability of a) rankings in Exp. 1a, b) BWS responses in Exp. 1b, before c) comparing elicited preferences across methods in Experiment 2.

Methods

To highlight the consistency across experiments, we describe all three experiments together.

Participants

Participants were recruited from the online platform Prolific Academic with the constraints that a) English was their first language, and b) had an approval rating greater than 80%, and c) participants did not participate in multiple experiments. Exp. 1a recruited 110 participants ($M_{age} = 41.7, SD = 12.9, N_{female} = 57,\ N_{male} = 52, N_{optout} = 1$), Exp. 1b recruited 101 participants ($M_{age} = 41.3, SD = 15.1,\ N_{female} = 55, N_{male} = 44, N_{optout} = 2$), and Exp. 2 recruited 161 participants ($M_{age} = 36.1, SD = 11.8,\ N_{female} = 66, N_{male} = 95$). From these, we removed 9, 1, & 13 participants respectively, where the experiment was only partially completed, where data did not save, or where participants took longer than the maximum allowable time to complete the study.

Participants were paid at a rate commensurate with £9.00 per hour ($\text{Med}_{time} = 6.75$ minutes). In total, these criteria left respective N's of Exp. 1a = 101, Exp. 1b = 100, and Exp. 2 = 148.

Design

All experiments followed a two-task structure where preferences for 10 life values were elicited twice. The experiments used a within-subjects (measurement occasion: first vs. second measurement) by between-subjects design (task labels: items vs. dimensions).

In Exp. 1a, preferences were twice elicited using the ranking task. In Exp. 1b, preferences were twice elicited using the BWS task. In Exp. 2, participants completed both ranking and BWS tasks, where the order of the tasks was randomised (i.e., ranking-first or BWS-first).

We manipulated the labels of the life values across both elicitation methods and measurement occasion. Values were labelled either as dimensions (e.g., hedonism) or example value-items (e.g., "Pleasure, enjoying life, self-indulgent", see Table 1 for full list of labels). These label sets were drawn from the Schwartz Value Survey (Schwartz, 1994) where previous work validated the items as representative examples of the 10 life values (Schwartz, 1992; Spini, 2003).

Label set (item vs. dimension) was randomised between-subjects. In Exp. 1a (ranking) and Exp. 1b (BWS), value labels were randomised such that $\frac{1}{3}$ participants completed both tasks with dimension-labels, $\frac{1}{3}$ with item-labels, and the remaining $\frac{1}{3}$ with label-changes across tasks; that is, we assume the order of completing the dimension- vs item-labels measure is inconsequential. In Exp. 2, we randomised labels across both task types and measurement occasions. That is, participants might see dimension labels *or* item labels at both measurement occasions, or they might see one of each label type across the two measurement occasions.

Materials

All 3 experiments elicited value preferences using two methods; a ranking task and a BWS task. Both tasks were conducted through the online survey platform QuestionPro and participants inputted their responses using a mouse.

Ranking task. In the ranking task, participants were presented with a randomised list of ten values labelled as either dimensions or items (see Table 1). To order the list, participants were instructed to drag individual life values into a separate field that automatically ordered the list from most (1) to least (10) important. Once all items were transferred to the ordered list, participants could proceed onto the next page.

BWS task. We adopted a 'case 1' object-type BWS task used by J. A. Lee et al. (2008) to examine responses about life values, using an exact replication of their stimulus set structure (see their Table 2). This stimulus set consisted of 11 subsets of Schwartz's life values, such as 'hedonism' and 'power', and asked participants to indicate the most- and least important life values within each subset (for example, see Figure 1).

Subsets consisted of lists of either 5 or 6 life values and, in total across the subsets, each life value was presented 6 times. The complete list of life values is presented in Table 1. Subsets were balanced such that pairs of values co-appeared on only three occasions. For example, 'hedonism' and 'power' appeared in the same subset on three occasions where the remaining values in each subset differed on each occasion. This randomisation is to ensure that each best-worst selection is weighed evenly against all remaining items shown in other subsets.

Participants were instructed to select the most- and least-important value within each subset before progressing to the next subset of values. Values were presented as a list with a 'most' indicator on the left of the list and a 'least' indicator on the right. There was no order restriction as to which preference was selected first.

Table 1

Dimension & item labels for all 10 life values ordered by averaged task scores in Figure 2. Participants were shown this table of definitions with both dimension & item labels at the outset of the experiment before completing the ranking or BWS choice task with only a single set of labels at any one time.

	Dimension labels	**Item labels**
1	Benevolence	Helpful, honest, forgiving
2	Self-direction	Creativity, curious, freedom
3	Achievement	Successful, capable, ambitious
4	Security	Clean, national security, social order
5	Universalism	Protecting the environment, a world of beauty, unity with nature
6	Hedonism	Pleasure, enjoying life, self-indulgent
7	Stimulation	Daring, a varied life, an exciting life
8	Tradition	Devout, accepting portion in life, humble
9	Conformity	Politeness, obedient, honouring parents & elders
10	Power	Social power, authority, wealth

Data processing & Outcomes

For the ranking task, we preserved the raw ranks of life values in data analysis. For the BWS task, best and worst choices were converted to Best-Worst scores (Louviere et al., 2015; Marley et al., 2016). Scores are generated by a count of each life value according to

$$\frac{N(most \text{ selections}) - N(least \text{ selections})}{N(appearances)}. \tag{1}$$

The calculated scores are bounded between -1 & 1 where values towards 1 indicate consistent 'most' important selections and values towards -1 indicate consistent 'least' important selections. For each participant, we calculated these normalised BWS scores for all 10 life values and produced an ordered list of values comparable to a rank ordered list.

Procedure

Participants began the online experiment by clicking a link to the experiment web page. At first, they were told the experiment was investigating life values and shown a table of values with both *dimension* and *item* labels, similar to Table 1. Participants were then randomly allocated into label conditions where value labels remained identical across both measurements (e.g., item-item or dimension-dimension) or value labels alternated (item-dimension or dimension-item).

Participants completed their first response task that was either a ranking task in Exp. 1a, a BWS task in Exp. 1b, or a random selection between the two tasks in Exp. 2. Once complete, participants proceeded onto the second task with the same task constraints. For Exp. 1a and 1b, this was the same task they completed in their first repsonse task, though dimension vs item labels may change. In Exp. 2, the second task was the task *not* completed first (i.e., if the ranking task was completed first then BWS was second, and if BWS was first then ranking was second). Participants were reimbursed upon completion.

Results

As an initial overview of the data, we first present an overall summary of participant preferences for the 10 Schwartz life values. Figure 2 panel A presents mean life value ranks from Exp. 1a and panel B presents mean BWS scores from Exp. 1b. Almost entirely, ranks and BWS scores were in agreement. For both tasks, the most important value was *benevolence* ($M_{rank} = 3.62, SE = 0.18$, $M_{BWS} = 0.375, se = 0.03$) and the least important value was *power* ($M_{rank} = 7.88, SE = 0.17$, $M_{BWS} = -0.46, se = 0.03$). Notably, the top five most important life values were identically ordered between ranking and BWS tasks as were the two least important life values. Only a single pair of values, *tradition* & *hedonism*, ranked 6th or 8th, traded ordinal positions between the preference elicitation methods.

As a reminder, across all three experiments, participants completed two preference elicitation tasks. In the next section, we present the agreement between both measurement

occasions using Kendall's Tau (τ). Kendall's Tau is a non-parametric statistic of rank correlation for ordinal data that provides a metric of similarity for lists of life-values generated by BWS as compared to ranking. Tau values are bounded between 1 & -1, where τ values of 1 indicate complete agreement across both lists, and τ values of -1 indicate exact reversals. For clarity, we use the τ_b variant of Kendall's Tau that explicitly accounts for ties between lists. Using the overview data from Figure 2 as an illustration, the order of mean ranks strongly agreed with mean BWS scores (items: $\tau = 0.96$, $BF_{10} = 221.4$, dimensions: $\tau = 0.82$, $BF_{10} = 43.0$, Figure 2).

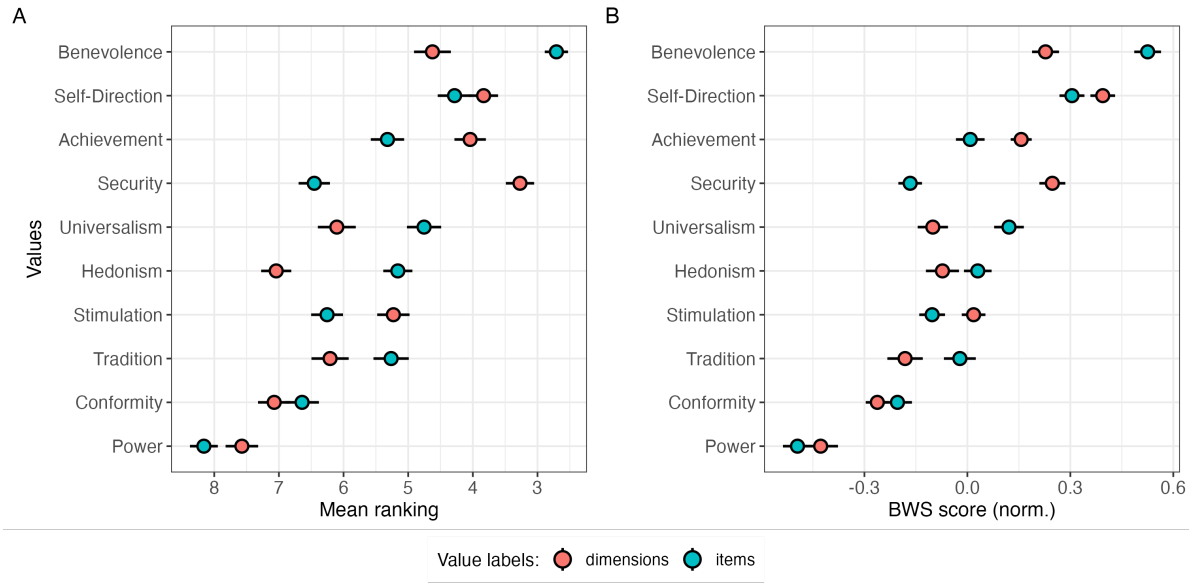

Figure 2. Overview of life value responses separated by task. Panel A shows mean ranks from Exp. 1A of each life value ordered from most important (1) to least important (10). Panel B shows mean BWS scores from Exp. 1B of each life value in the same order from most important (1) to least important (-1). Value label is shown in colour and error bars represent standard error of the mean.

Experiment 1a (ranking) & 1b (BWS)

We first examine the preferences across repeated-measurement occasions within the same task (Exp. 1a & 1b) and then across elicitation tasks (Exp. 2), shown in Figure 3. Note

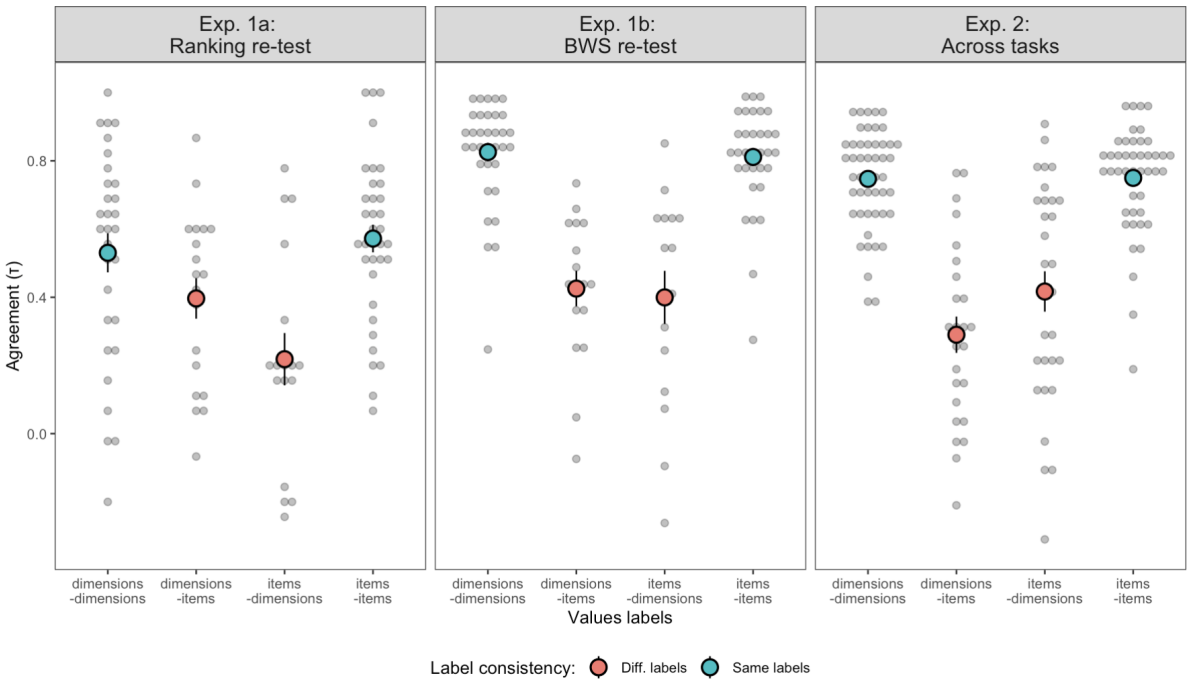

Figure 3. Agreement across measurement occasions as a function of life value labels (x-axis) and experiment (panels). Gray dots show individual-level participant data and larger coloured dots show means and standard error. Colour indicates life value label consistency. Blue coloured dots indicate conditions where life value labels were the same on both measurement occasions, and red colour dots indicate when labels differed. Exact ordering of labels is shown on the x-axis. See Table 1 for complete list of life-value labels.

that in the following sections, agreement between measurements is indexed by Kendall's Tau (τ) which is shown at the individual participant-level within the Figure, and reported in the aggregate in-text.

Beginning with Exp. 1a & 1b, the agreement across first and second measurements was higher, on average, for the BWS task compared to ranking task (i.e., the average of individual participant Kendall's Tau values shown in the left-most vs. centre panel Figure 3, $M_{rank} = 0.47$ vs. $M_{BWS} = 0.69$, *SE's*= 0.03, $BF > 1000$). This advantage for BWS tasks likely emerges from the relative ease of selecting preferences at the extremes. In line with this interpretation, the most common *rank* to change across measurement occasion in

Exp. 1a was 6, one of the middle ranks (81% of rank 6 responses changed) and the consistency of ranked values across measurement occasions only increased for each rank-step toward either extreme. That is, by comparison, rank 1 changed for only 54% of responses, rank 10 responses changed for 62%, and other rank-changes monotonically increased with each rank-step before peaking at rank 6.

Across both Exp. 1a & 1b, agreement was stronger with consistent life-value labels than inconsistent labels (blue vs. red points, Figure 3, $M_{1a} = 0.55$ vs. 0.31, $M_{1b} = 0.82$ vs. 0.41, BF > 1000). That is, changing the label of the life values between measurement occasions decreased the agreement between surveys (e.g., benevolence $\leftrightarrow$ helpful, honest, forgiving).

Notionally, agreement should differ if the new label evokes qualitatively different definitions of a particular life value. This divergence in people's interpretations of item and dimension labels appears in the aggregate preference, shown in Figure 2. Despite the divergence, however, one interesting consistency is that across both ranking and BWS tasks, the relative preference for an item or dimension label for any *individual life value* was the same (red vs. blue points across panels in Figure 2). From a construct validity perspective, this consistency is reassuring as it indicates that despite changing elicitation method we indexed the same underlying preferences for each label set, at least in the aggregate. We return to this point in the General Discussion.

One point worth noting is that the BWS task allows for ties in preference where ranks do not. These ties indicate equivalent (non-)preferences between life values and so for 10 life values, there may be fewer than 10 bins of BWS scores. In Exp. 1b, where individuals completed the BWS task twice, the mean number of BWS-score bins were similar across the first measurement occasion and the second measurement ($M = 7.06$ vs. 6.89, $BF_{01} = 2.82$). Notably, both occasions exhibited fewer than 10 bins indicating the presence of ties in preferences where ranks in Exp. 1A may have forced discrimination.

Experiment 2 (crossed-tasks)

In Experiment 2, we directly compared preference elicitation methods such that participants completed both BWS and ranking tasks. Overall, preferences elicited by BWS methods largely agreed with preferences elicited by ranking (right-most panel, Figure 3, $M_\tau = 0.60, SE = 0.02$). However, similar to Exp. 1a and 1b, consistent labeling led to greater agreement across ranking and BWS methods (blue vs. red points, $M_\tau = 0.75$ vs. 0.36, $BF = 9.19$). To understand better where rankings and BWS specifically diverged, we now turn to examine raw BWS scores from Experiment 2.

Figure 4 shows the mean BWS scores for each rank position as a function of label consistency. Although both measures largely agree on ordinal terms, the relative distances between ranks may guide where disagreement is likely to emerge. For conditions with consistent labeling, one immediate observation is that BWS scores for the first and last rank are more distant from their neighbouring ranks compared to any other rank positions ($\Delta_{1-2} = 0.40, \Delta_{9-10} = 0.38$ vs. $M_\Delta = 0.11, se = 0.01$). This discrepancy suggests that preferences for first and last ranked values are generally stable. By comparison, average BWS scores for middle ranks, such as the fourth, fifth and sixth ranks, are more tightly clustered suggesting that preferences for options in these rank positions are less discernible from one another. The conflation of similar BWS scores is particularly exaggerated for conditions where the labels switched between tasks (right-most x-axis points, Figure 4). Broadly, the distribution of ranks is more centrally concentrated and within each rank position, each mean BWS score is more variable compared to conditions with the same labels. Considered together, mean BWS scores across rank positions provide a tentative guide as to where disagreement between ranks and BWS arises.

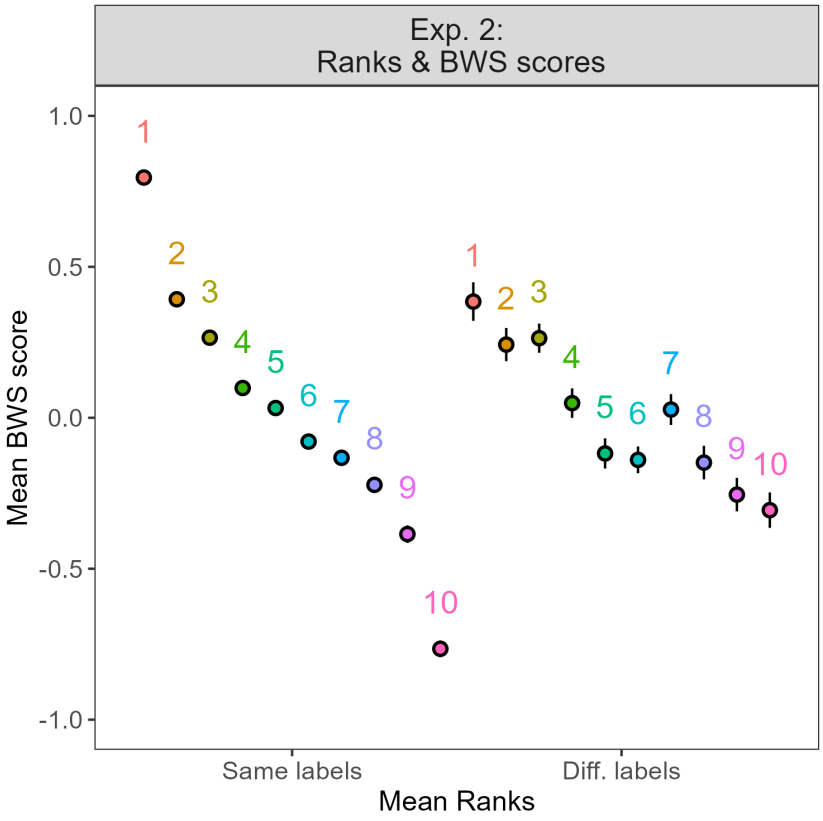

Figure 4. Experiment 2 aggregate-level data showing mean BWS scores (y-axis) as a function of ranks (x-axis). For each rank position, the mean BWS score across individuals is shown in the coloured points. Error bars represent standard error of the mean and text indicates corresponding rank positions. Between-subject label conditions are separated on the x-axis. As a reminder, *different labels* indicate dimension-item labels switched with ranking and BWS tasks whereas *same labels* indicate consistent labeling across tasks.

General Discussion

In this paper, we contrasted two popular methods of preference elicitation. We compared rankings, which extract ordinal preferences, to BWS methods that additionally provide relative information about the strength of those preferences. Across three experiments, we examined the agreement between elicited preferences over time in Exp. 1a and 1b, across elicitation methods in Exp. 2, and over changes in item labels. Together, we draw two

main conclusions.

First, preferences elicited from BWS and ranking agree with one another. Despite the differences in response methods and task demand on individuals, we find that elicited preferences were largely consistent indicating that both methods captured a common underlying representation of individual's preferences about their life values. For survey-designers, our findings provide an empirical validation to support using BWS in contexts where ranking methods are currently employed. The primary benefit is that not only are ordinal preferences between methods preserved but BWS also provides valuable relative information between preferences and detects ties in preference where rankings cannot.

Second, BWS has greater test-retest reliability compared to rankings. A core methodological consideration is whether one's measurements are stable over time. From Experiment 1a, we find that while acquiring a set of rankings is relatively inexpensive in time-costs, not all ranks are valued equally. Specifically, test-retest reliability across sets of rankings is harmed by interchanging middle ranks that are generally less discernible compared to preferences at the extremes. BWS capitalises on this exact psychological feature. We found that the cognitive ease of identifying the best and worst options in BWS produces higher test-retest reliability across measurement occasions.

Taken together, our results privilege BWS over rankings as it provides a richer characterisation of preferences alongside the reassurance of test-retest reliability.

Different labels, same interpretations?

Our work leaned on the conceptual similarities between items and dimensions. Across experiments, we labelled each life value using either a dimension label (e.g., benevolence) or an item label (e.g., helpful, honest, forgiving). These labels were drawn from labels sets within Schwartz' Value Survey (Schwartz, 1994) where previous work conducted across 21 countries had validated the set of items as representative examples of the 10 life values

(Schwartz, 1992; Spini, 2003).

When thinking about life values, however, context provides many cues to guide interpretation. For instance, the personal importance of the life value 'power' is likely to differ if one individual defines power as social influence while another individual defines power as a proxy for wealth. Alongside individual differences, however, one important consideration is whether the elicitation method itself led to divergent definitions. Should one method inherently guide individuals towards a particular definition, then agreement between the methods would be systematically compromised by the item set rather than the method. We addressed this possibility by incorporating two sets of item labels and comparing the preferences extracted from each.

Across experiments, item and dimension label sets led to different interpretations of the life values. Where this is most apparent is that in all three experiments, agreement was lower for conditions where label sets alternated. Notably, this disparity held for twice-attempted ranking tasks in Exp. 1a and BWS in Exp. 1b suggesting that different interpretations were present independent of the elicitation method.

Despite the differences *between* the label sets, one interesting consistency is that their definitions appear to have been shared *across* elicitation methods. This concurrence is evident in the relative preference data. In Figure 2, the order of preference for either item or dimension labels is consistent across ranks and BWS scores. For example, in both ranking and BWS methods, 'benevolence' was more preferred when labelled as items 'helpful, honest, forgiving' compared to as a dimension. Similarly, 'security' was more preferred when labelled as a dimension rather as items 'clean, national security, social order'. This likely emerged because without the context to imply a societal definition of security, many individuals may have converged on a common alternative definition such as personal security. In addition to these discrete label preferences, the relative magnitude of preference changes between item-labelled and dimension-labelled is also similar across elicitation methods. Considered together, these two patterns are consistent with the idea

that individuals converged on distinct definitions for items and dimensions.

Considering set size

Broadly, the consistency in preferences across elicitation method speak strongly against set-size context effects. Although BWS methods only present a subset of the 10 life values, the similarity in elicited preferences in the aggregate (Exp. 1a & 1b, Figure 2) and at the individual-level (Exp. 2, Figure 3) suggests that choosing from a smaller subset is not a diminished compromise of choosing from a complete list.

Compared to rankings, a core strength of BWS is that preferences across larger sets are as easily obtained as preferences for smaller sets. For only 10 life values, our experiments found that rankings in the middle order were more likely to change across measurement occasion. Hypothetically, one might expect that increasing the set size would only exaggerate the paucity of information from middle ranked values while adding greater cognitive effort of comparing each item against a larger set of alternatives. When one considers the large number of possible influences on surgeon choice ($N = 16$ in Ejaz et al. (2014)) or the responses in personality surveys ($N = 57$ in Schwartz (1994)) it becomes clear how ranking each option relative to a full set is problematic.

BWS sidesteps this set size problem due to its iterative piecemeal design. The main mechanism that permits this is that for any set size, BWS methods require only two choices from a subset of all items. For example, our experiments presented subsets of 5 or 6 life values, from which only the best and worst is chosen. This feature means larger set sizes only increase the number of choice 'rounds' an individual completes while maintaining the same cognitive load demand across each round.

Naturally, this advantage is not free. Increasing the number of rounds means that for the respondent, BWS surveys can take longer to complete as compared to equivalent set-size ranking surveys, as was borne out in our completion time data (Exp. 1a vs. 1b, $M_{rank} = 4.60$ vs. $M_{BWS} = 8.60$ minutes). However, our findings suggest that the added

reliability afforded by the time-cost is worthwhile considering.

Taking a broader lens, our work dovetails with the long-standing tradition of examining preferences in the absence of ground truth. Like with music taste or choosing between medical treatment options, understanding the subjectivity is the objective. One notable and recent exception is that of Gronau et al. (2023) whereby classic likert elicitation methods and choice tasks are compared in settings where a 'truth' was engineered, albeit modifying the standard preference elicitation use case. Akin to their efforts, the goal of this work was to evaluate two popular elicitation methods but by standards of consistency, preserving their core task qualities - subjective warts and all. By these standards, we arrive at the conclusion that the BWS method is more reliable than ranking.

Declarations

Funding

This work was supported by the joint MURI-AUSMURI in Cybersecurity Assurance for Teams of Computers and Humans (CATCH) and funding from the Australian Research Council (DP210100313).

Conflicts of interest

The authors have no conflicts of interest to declare.

Ethics approval

This research was approved by the Human Research Ethics Committee at the University of Newcastle, Australia under protocol number H-2019-0321.

Consent to participate

All participants consented to participate in the experiment.

Consent to publish

All authors consent to publishing this manuscript.

Availability of data and materials

Survey materials were collected on the QuestionPro platform and can be shared on request. Raw and preprocessed data are available on OSF.
https://osf.io/ct8ae/?view_only=5dfea3dcf2e14bc3a6f360562883cab0

Code availability

Analysis & data preparation code are available at the above OSF repository.

References

Cooper, G. J., & Hawkins, G. E. (2019). Investigating consumer decision strategies with systems factorial technology. *Journal of Mathematical Psychology*, *92*, 102258.

Edwards, W. (1954). The theory of decision making. *Psychological bulletin*, *51*(4), 380.

Ejaz, A., Spolverato, G., Bridges, J. F., Amini, N., Kim, Y., & Pawlik, T. M. (2014). Choosing a cancer surgeon: Analyzing factors in patient decision making using a best–worst scaling methodology. *Annals of surgical oncology*, *21*, 3732–3738.

Grimmond, J., Brown, S. D., & Hawkins, G. E. (in press). A solution to the pervasive problem of response bias in self reports. *Proceedings of the National Academy of Sciences*.

Gronau, Q. F., Bennett, M. S., Brown, S. D., Hawkins, G. E., & Eidels, A. (2023). Do choice tasks and rating scales elicit the same judgments? *Journal of choice modelling*, *49*, 100437.

Isaac, M. S., & Schindler, R. M. (2014). The top-ten effect: Consumers' subjective categorization of ranked lists. *Journal of Consumer Research*, *40*(6), 1181–1202.

Lee, J. A., Soutar, G., & Louviere, J. (2008). The best–worst scaling approach: An alternative to schwartz's values survey. *Journal of personality assessment*, *90*(4), 335–347.

Lee, M. D., Steyvers, M., De Young, M., & Miller, B. (2012). Inferring expertise in knowledge and prediction ranking tasks. *Topics in cognitive science*, *4*(1), 151–163.

Louviere, J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Louviere, J., Lings, I., Islam, T., Gudergan, S., & Flynn, T. (2013). An introduction to the application of (case 1) best–worst scaling in marketing research. *International journal of research in marketing*, *30*(3), 292–303.

Marley, A., Islam, T., & Hawkins, G. E. (2016). A formal and empirical comparison of two score measures for best–worst scaling. *Journal of Choice Modelling*, *21*, 15–24.

Marley, A., & Louviere, J. (2005). Some probabilistic models of best, worst, and best–worst choices. *Journal of mathematical psychology*, *49*(6), 464–480.

Montgomery, L. E., Bradford, N., & Lee, M. D. (2024). The wisdom of the crowd with partial rankings: A bayesian approach implementing the thurstone model in jags. *Behavior Research Methods*, *56*(7), 8091–8104.

Rokeach, M. (1973). The nature of human values. *New York: The Free Press*.

Ryan, M. (2004). Discrete choice experiments in health care. *Bmj*, *328*(7436), 360–361.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology/Academic Press*.

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of social issues*, *50*(4), 19–45.

Spini, D. (2003). Measurement equivalence of 10 value types from the schwartz value survey across 21 countries. *Journal of cross-cultural psychology*, *34*(1), 3–23.

Van Dam, A. (2024, May). *America's best decade, according to data*. https://www.washingtonpost.com/business/2024/05/24/when-america-was-great-according-data/

Viney, R., Lancsar, E., & Louviere, J. (2002). Discrete choice experiments to measure consumer preferences for health and healthcare. *Expert review of pharmacoeconomics & outcomes research*, *2*(4), 319–326.

Von Neumann, J., & Morgenstern, O. (2007). Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press.