Journal of Choice Modelling 30 (2019) 61–72

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: www.elsevier.com/locate/jocm

Best-Worst Scaling with many items

Keith Chrzan[a,*], Megan Peitz[b]

[a] Sawtooth Software, Inc., 996N 250E, Chesterton, IN, 46304, USA
[b] Sawtooth Software, 3210 N. Canyon Road, Provo, UT, 84604, USA

ARTICLE INFO

Keywords:
Best-worst scaling
Discrete choice experiments
Large numbers of items

ABSTRACT

Best-worst scaling (BWS) has become so useful that practitioners feel pressure to include ever more items in their experiments. Researchers wanting more items and enough observations of each item by each respondent to support individual respondent-level utility models may greatly increase the burden on respondents, resulting in respondent fatigue and potentially in lower quality responses. Wirth and Wolfrath (2012) proposed two methods for creating BWS designs that allow for large numbers of items and respondent-level utility estimation, Sparse and Express BWS. This study aims to uncover the recommended approach when the goal is recovering individual respondent-level utilities and intends to do so by comparing the relative ability of Sparse and Express BWS to capture the utilities that would have resulted from a full BWS experiment, one with at least three observations of each item by each respondent. The current study repeats previous comparisons of Sparse and Express BWS using a new empirical data set. It also extends previous findings by collecting enough observations from each respondent for both a full experiment and one of the proposed methods, Express BWS and Sparse BWS. The results replicate and extend previous findings regarding the superior ability of the Sparse BWS methodology, relative to Express, to reproduce "known" utilities or utilities that result from a full BWS design.

1. Introduction

1.1. Best-Worst Scaling

Rating scales suffer from significant shortcomings: different respondents use them differently, producing response set effects or scale use bias (Baumgartner and Steenkamp, 2001). A well-known halo effect (Thorndike, 1920) can cause artificially high correlations among ratings of disparate constructs, even when the meaning of those constructs would suggest zero or even negative correlations (in Thorndike's memorable phrasing, the correlations he found among logically unrelated items were "too high and too even"). Finn and Louviere (1992) developed Best-Worst Scaling (BWS) as an alternative to rating scales, designed to elicit powerful measures of relative preference using respondent-friendly questions. For example, in the study of 36 dessert items described in Section 3, the BWS experiment includes a series of questions listing four desserts each; each question forces the respondent to make tradeoffs in that she can choose a single most liked dessert and a single least liked dessert (see Fig. 1 below).

Three variants of BWS now exist. BWS Case 1, the "object" case, involves respondent evaluations of best and worst items or objects from short lists of items. Case 1 is the form of BWS first introduced by Finn and Louviere (1992) and it features designed choice sets presented in a straightforward respondent task which, when subjected to appropriate analysis, enables a researcher to place multiple items on a single scale (Marley and Louviere, 2005; Louviere et al., 2015). BWS Case 2, the "profile" case, has

* Corresponding author.
E-mail address: keith@sawtoothsoftware.com (K. Chrzan).

https://doi.org/10.1016/j.jocm.2019.01.002
Received 4 August 2017; Received in revised form 1 January 2019; Accepted 6 January 2019
Available online 11 January 2019
1755-5345/

Now imagine that you've had an evening meal at the casual dining restaurant and that you've decided to buy a dessert. Considering only these 4 desserts, which would you like the Most and which would you like the Least?

(1 of 27)

Most		Least
○	Coconut cream pie	○
○	Mango lassi	○
○	Crème brûlée	○
○	Vanilla milk shake	○

Fig. 1. Example BWS question.

respondents choose best and worst attribute levels from conjoint-style profiles, while Case 3, the "multi-profile" case, has respondents choose best and worse profiles from choices sets of multiple profiles (Louviere et al., 2015). This paper concerns only BWS Case 1, henceforth BWS for short. As a general scaling method, BWS can be used, for example, to measure the relative appeal of public policy options or of new product ideas (Finn and Louviere, 1992); or the relative importance of product attributes in a purchase decision (Lockshin and Cohen, 2015); or the extent to which respondents agree with items on a psychometric scale (Lee et al., 2007).

The set of BWS questions each respondent sees conform to an experimental design, which may be common across all respondents, unique for each respondent or blocked so that different groups of respondents receive different sets of questions. Early versions of BWS used orthogonal main effects designs but today researchers use balanced incomplete block designs (when they exist) or designs found through computer search algorithms (Kuhfeld and Wurst, 2012).

As for any other successful methodology, end users of BWS results press researchers to do more: more studies, across more subject areas, and, especially, with more items. We note that this problem does not seem to affect many published papers on BWS – the book by Louviere, Flynn and Marley includes case studies with only up to 13 items, while papers from the past four years of *The Journal of Choice Modeling* had as many as 17 items. Practitioners, however, frequently study many more items: in the past year nearly a third of the authors' studies have involved 30 or more items, nearly one in 10 involves 50 or more items and a handful involve 100 or more items. The authors are also aware of a client doing several studies a month with 100 + items.

1.2. Large numbers of items

But as a BWS experiment includes more items, the demands on the respondent increase. In a standard best-worst experiment where we want to show each item three times, one can determine the appropriate number of questions, depending on the total number of items using simple math, multiplying three times the number of items in the study and dividing by the number of items per question. Increasing the total number of items increases the number of questions per respondent, and thus the length of the survey, at some point adversely affecting the quality of responses.

There are two methods developed to accommodate large numbers of items, namely Sparse and Express BWS, described in section 2. This paper reviews prior research comparing the two methods, in which Sparse BWS appears to have an edge over Express BWS. We then combine the strengths of the previous comparative studies in a new empirical study, the results of which point even more strongly to Sparse BWS as the better method to handle large numbers of items.

As noted, marketing research practitioners have clients who want to include large numbers of items in BWS surveys, often more than 30 items and sometimes many more than that. In addition to having large numbers of items, applied researchers often need individual respondent-level BWS utilities for use in subsequent analyses like factor analysis, reliability analysis, cluster analysis and TURF (Total Unduplicated Reach and Frequency, a method for prioritizing bundles of items borrowed from the world of advertising research). The simultaneous pressure of having many items and requiring high-quality respondent-level utilities means that the standard advice that each respondent sees each item three or four times will result in a very lengthy survey (e.g. 75 items shown in sets of 5 items each would require 45 BWS questions per respondent in order for each respondent to see each item three times). To some extent, this resembles the problem of too many attributes in stated choice experiments for which several solutions have also been proposed (Green et al., 1981; Chrzan, 2010; Zhang et al., 2015). Again, however, this problem appears to be one that affects practitioners and one that has not heretofore appeared in the academic literature. Several methods have been proposed in the practitioner literature to reduce the respondent burden.

2. Research background

2.1. Modeling best-worst choices

Assuming the best and worst choices derive from a common underlying utility function, one combines the best and worst choices into a single model that results in a separate score, or utility, for each item. Perfectly balanced BWS experiments (e.g. those using balanced incomplete block designs) may be analyzed easily, without recourse even to computer analysis, simply by subtracting the

number of times each item is identified as worst from the number of times it is selected as best and comparing these net counts across the items (Louviere et al., 2015). Lipovetsky and Conklin (2014) offer an additional utility estimation formula which Marley et al. (2016) found to predict in- and out-of-sample choices better than another common count-based estimator of BWS utilities. One can also use statistical models to extend the ability to estimate utilities to cases of imperfectly balanced experimental designs: aggregate multinomial logit, MNL, for sample-level utility estimates (McFadden, 1974) or latent class MNL (DeSarbo et al., 1995) for segment-level utility estimates. In the field of marketing, interest in developing products for market segments has led to the widespread adoption of hierarchical Bayesian mixed logit (Allenby and Ginter, 1995) which produces individual respondent-level utility estimates. In a study using artificial respondents with known utilities, Orme (2005) showed that using choice sets of four to five items and enough sets to enable each respondent to see each item in three to four sets suffices for reliable respondent-level utility estimation in BWS studies. Subsequently, we refer to an experiment in which each item appears in at least three choice sets per respondent as a "Full" BWS experiment.

2.2. Sparse and express BWS

Two approaches for accommodating large numbers of items in a BWS experiment come from Wirth and Wolfrath (2012) who propose methods that allow for potentially many more items than standard BWS (the authors are aware of commercial studies of 100 or more items that have used each of these two methods). Express BWS shows different respondents different subsets of the items; each respondent sees each of the items in her subset (and only those items) three or four times. For example, in a study of 100 items, we might randomly select a unique subset of 24 items for each respondent. Each respondent receives 18 choice sets of four items each, with each of the 24 items in her unique subset presented in three choice sets, and with none of the other 76 items presented. Each respondent receives a different subsets of items, but across respondents, every item would be included in some respondents' subsets. A Sparse BWS, on the other hand, allows each item to appear as infrequently as just once across each respondent's BWS questions (rather than the previously suggested three to four times). To continue the example study of 100 items, we might, with a Sparse BWS experiment, show each respondent 20 sets of five items each. A given respondent would see all 100 items, with each item appearing in exactly one choice set. Different respondents would receive different blocks of the experimental design, allowing each item to appear with all the other items as well as enabling us to control for order and position effects.

2.3. Previous comparisons of sparse and express BWS

Wirth and Wolfrath (2012) report that Sparse BWS performs slightly better than Express BWS in terms of predicting best and worst choices in holdout choice sets in an empirical study. They also report the results of a Monte Carlo study of Express BWS that shows it recovering the known parameters of artificial respondents well – almost perfectly for sample-level utilities and reasonably well for respondent-level utilities. They did not, however, conduct a similar analysis for Sparse BWS.

Chrzan (2015) expanded upon Wirth and Wolfrath's artificial data analysis. His study compared Sparse and Express BWS using two artificial data studies based on individual respondent-level utilities from human respondents in commercial research. In both studies, Sparse and Express BWS recovered known mean sample-level utilities accurately: correlations averaged 0.995 across the two studies for Sparse and 0.989 for Express, a statistically significant but practically ignorable difference. The two methods differed in their ability to reproduce utilities at the individual respondent-level, however, with Sparse at a 0.804 correlation with known utilities versus 0.743 for Express, a statistically significant and meaningfully large difference between the two methods.

Serpetti et al. (2016) conducted further investigation among human respondents. Using 2,202 respondents, Serpetti et al. showed that in-sample, Sparse was better able than Express to mimic item's rankings from a Full BWS in-sample (correlation of 0.49 versus 0.41) and out-of-sample that Sparse produced lower MAE than Express (0.034 versus 0.038).

Thus, findings to date agree that Sparse BWS handles large item sets better than Express BWS. However, previous studies have either compared Sparse BWS and Express BWS in terms of how well their sample mean utilities compare to the mean utilities from a Full BWS experiment; or they used artificial respondents to investigate the ability of Sparse and Express BWS to reproduce known respondent-level utilities. We seek, with the current study, to repeat previous analyses conducted at the level of mean utilities for samples of respondents while also expanding our analysis to compare individual respondent-level utilities estimated from Sparse and Express BWS to individual respondent-level utilities from a Full BWS experiment, for those same human respondents. In other words, we seek to extend the Chrzan (2015) respondent-level parameter recovery experiment to human respondents, treating respondents' Full BWS utilities as the "known" utilities we want to be able to reproduce. This approach will provide another comparison of whether sample mean utilities for Sparse or Express BWS better reflect those from a Full BWS. Extending the analysis to assess the ability of Sparse and Express BWS to recover known individual respondent-level utilities (from a Full BWS experiment) adds a powerful new comparison not previously applied to human respondents.

3. Current empirical study

3.1. Methods

3.1.1. Research design

In a survey of 1,207 recent customers of casual dining restaurants, we asked respondents about their relative preference for 36 dessert items (Table 1). (See Fig. 1 for an example Best-Worst question)

Table 1
All 36 dessert items.

Item #	Item
1	Crème brûlée
2	Italian gelato
3	Belgian waffle ice cream cone
4	Churros
5	French silk pie
6	Cinnamon roll
7	Coconut creme pie
8	Apple pie
9	Chocolate molten cake
10	Mango lassi
11	Chocolate mousse
12	Bread pudding
13	Hot fudge sundae
14	Salted caramel sundae
15	Vanilla milk shake
16	Strawberry milk shake
17	Chocolate layer cake
18	Lemon meringue pie
19	Flan
20	Skillet chocolate chip cookie
21	Cherry pie
22	Texas pecan pie
23	Louisiana mud pie
24	Peach cobbler
25	Cinnamon crumble cake
26	New York cheesecake
27	Oreo cheesecake
28	Peanut butter cheesecake
29	Hot brownie sundae
30	Rice pudding
31	Key lime pie
32	Tiramisu
33	Red velvet cake
34	Blueberry tart
35	Cannoli
36	Pumpkin cheesecake

The respondent in this example identifies the one dessert she would *most* enjoy and the one dessert she would *least* enjoy, among the set of four. Subsequent questions would include different subsets of four of the 36 desserts until the respondent ended up seeing each dessert a few times.

Because our primary contribution in this paper involves comparing the ability of Sparse and Express BWS to reproduce known utilities at the individual respondent-level, and because we know those utilities only by estimating a Full BWS experiment for those respondents, we wanted to limit the number of items to what we could confidently believe a Full BWS experiment could handle. Extending to 100 or 200 items would have prevented us from relying on the utilities from the Full BWS (indeed, doing so would be an instance of the problem researchers use Sparse and Express BWS to solve).

For purposes of this research, we randomly assigned each respondent to one of three cells: the Sparse BWS treatment cell, the Express BWS treatment cell and a Full BWS control cell (See Table 2).

For the Full BWS control cell, each respondent received one of 100 blocks of 27 sets of quads, where each block results from a computer search for an efficient experiment that balances the frequency with which each item appears, orthogonality and the frequency with which each item appears in each of the four positions in the choice set (top, second, third, bottom). Having 100 blocks allows us to cover a larger part of the design space than would a single version, and it allows us to even out order and position effects as well. The 27 quads allow each respondent to see each of the 36 dessert items three times, making this a Full BWS design. A total of

Table 2
Comparison of full, sparse and express experimental designs.

Research Design for Comparison of Sparse and Express BWS to Full BWS Experiment	Full BWS	Sparse BWS	Express BWS
# of Choice Sets	27	9	9
# of Items/Choice Set	4	4	4
# of Items to be tested per respondent	36	36	12
Average # of times each item is seen	3	1	3
Total N	403	400	404

Table 3
Additional detail on sparse experimental design.

Research Design for Comparing Sparse BWS with Full BWS on the same sample	Sparse BWS	Additional 18 tasks given to Sparse Cell	Full Sparse BWS
# of Choice Sets	9	18	27
# of Items/Choice Set	4	4	4
# of Items to be tested per respondent	36	36	36
Average # of times each item is seen	1	2	3

403 respondents completed the Full BWS.

The Sparse BWS treatment cell contained 400 respondents. We made a single, highly efficient block for the design, again looking to balance the frequency with which each item appears (once per respondent in this case) and the frequency with which each item appears in each position. For each respondent, we randomized the assignment of desserts to design positions. Thus the dessert attached to the 26th item in the design might be the New York Cheesecake for one respondent and Texas Pecan Pie for the next. In effect, each respondent received a unique set of experimental stimuli. Respondents in this cell received nine sets of quads, with each item appearing just once. We also asked each respondent two more efficient sets of nine quads each, so that each respondent still sees 27 questions, resulting in each item being seen three times. Showing each item in 27 choice sets allows us to estimate Full BWS results even for respondents in the Sparse BWS treatment cell, thus enabling us to test the ability of Sparse BWS to recover the known, individual respondent-level utilities of a Full BWS experiment (See Table 3).

The Express BWS treatment cell had 404 respondents. The first nine quads of each respondent's BWS experiment includes an individualized random subset of 12 dessert items out of the full set of 36 dessert items. In those first nine quads, each respondent sees each item in his random subset of 12 dessert items three times. Utilities from only these nine quads feature in the direct comparisons of the estimated Sparse and Express BWS models. As in the Sparse BWS treatment cell, we also asked each respondent in this cell 18 more efficiently designed quads, with each of the 36 items appearing twice each. Therefore, on average, across the full 27 sets of quads, each respondent sees each item three times (12 of the items five times each and the other 24 items just twice). We used computerized search software to identify efficient designs, in 100 blocks each, for both the first nine and the final 18 choice sets in the Express BWS cell, thus controlling for order and position effects in the Express BWS treatment cell (See Table 4).

3.1.2. Model estimation

To create BWS scores, we assume respondents consider all k items in the set and choose one pair that maximizes the distance between the ‘best’ item and the ‘worst’ item (this is also why BWS is often referred to as MaxDiff). Because we have best and worst responses, the independent variable matrices are dummy-coded as two separate sets: one for best and one for worst responses. If a design matrix X describes best choices then design matrix -X describes worst choices and we concatenate both designs in utility estimation (pooled estimation). Bacon et al. (2007) compared this estimation method with one that codes best-worst choices to identify maximally differing items in choice sets and concluded that utilities produced by the different methods “did not systematically differ, despite our concerted effort to show the contrary.” Dyachenko et al. (2014) suggest a sequential estimation of BWS but pooling of best and worst choices seems to be the most common method for estimating BWS utilities, used in both the analytical Best-Worst estimator (Lipovetsky and Conklin, 2014) and the difference in best and worst counts described by Louviere et al. (2015), as well as in commercial BWS software (Sawtooth Software, 2013).

Each of the five utility sets (Full, Sparse, Express, Sparse Full, Express Full) will be estimated using a hierarchical Bayesian (HB) mixed logit model (Allenby and Ginter, 1995). This “random-effects" model assumes that the respondent weights are normally distributed in the population. To avoid linear dependency in the model, the last level is omitted and the utility is constrained to zero. All other K-1 levels are estimated with respect to that level's zero parameter. In this study, Pumpkin Cheesecake, item 36, was the omitted level.

HB borrows information across the sample to stabilize the estimates for each individual. It is called “hierarchical” because it has two levels. At the higher level, the individuals' part-worth utilities are assumed to be described by a multivariate normal distribution, characterized by a vector of means and a matrix of covariances. At the lower level we assume that the probability of an individual choosing a particular alternative, given that individual's part-worths, is governed by a multinomial logit model (MNL).

For each of the five utility models, we planned to use 20,000 burn-in iterations and 10,000 saved iterations. Burn-in iterations are done before convergence is assumed and are not saved. The saved iterations are used in analysis to develop the point estimates. The final point estimate is an average of the saved iterations for each respondent.

Table 4
Additional detail on express experimental design.

Research Design for Comparing Express BWS with Full BWS on the same sample	Express BWS	Additional 18 tasks given to Express Cell	Full Express BWS
# of Choice Sets	9	18	27
# of Items/Choice Set	4	4	4
# of Items to be tested per respondent	12	36	36
Average # of times each item is seen	3	2	3

One must also declare the Degrees of Freedom (5) and Prior Variance (1.0) prior to model estimation. The prior degrees of freedom refer to the covariance matrix and do not include the number of parameters to be estimated. A prior variance reflects the weight on fitting each individual's data, versus the amount of information borrowed from the population parameters.

The model produces raw logit-scaled parameters for each individual that we then zero-centered and transformed to a 0–100 scale. These individual respondent-level results are then averaged to create a sample utility measure. Those sample utility measures were then transformed into a ranking of the 36 items. This resulted in both a sample utility measure and a sample rank for 36 desserts for each experiment. The sample utility measure, standard deviation and ranking are presented in Appendix 6 for each of the five models. The methods of rescaling are also reported in Appendix 6.

3.1.3. Planned comparisons

First, we want to test whether Sparse or Express BWS does a better job of predicting the results of the control cell (the Full BWS experiment). For this comparison, we will estimate the mean BWS utilities for each of the 36 desserts according to the model estimation plans in section 3.1.2, in each the Full BWS experiment (27 quads), Sparse BWS (nine quads) and Express BWS (nine quads). Then we will average the individual respondent-level utilities to get one sample utility score for each of the 36 desserts, for each of the three experiments. Next, we will compare the sample utilities in the Sparse and Express BWS experiments with the sample utilities in the control cell (the Full BWS experiment) using the test of dependent correlations (Cohen and Cohen, 1983). Applied researchers sometimes report ranks rather than BWS utilities, so while we expect similar results, we will also look at the rank ordering of the sample utilities of the 36 desserts for all three experiments. We will compare the Sparse BWS sample rankings and Express BWS sample rankings with the control cell sample rankings, again using the test of dependent correlations.

We also plan to assess the ability of Sparse and Express BWS to recover “known” parameters at the individual respondent-level. For this analysis, we assume that the estimation from each respondent's set of 27 quads makes for a more valid measure of that respondent's utilities than would estimation from the smaller set of nine quads, be they Sparse or Express BWS, thus we call them “known” parameters. For this analysis, we will use the individual respondent-level utility estimates for the nine quads for the Sparse and Express BWS Experiments already estimated. We employ the same model estimation settings as described in Section 3.1.2 for the respondents' 27 sets of quads for each the Sparse and Express BWS experiments. These utility estimates are what we will refer to as the “known” utility estimates for both experiments. Two sets of utilities now exist for the 36 dessert items per respondent - one for the nine sets of quads and the other for the complete 27 sets of quads, dependent upon which exercise they completed – the Sparse BWS or Express BWS. For example, each of the 400 respondents in the Sparse BWS cell provides two sets of utilities for each of the 36 dessert items, one set “known” and one set Sparse. We concatenate these 36 utility estimates (nine quads, Sparse) and “known” (27 quads) utility estimates across all 400 respondents so that we have a 2 by 14,400 matrix for each of the two treatment cells' correlations. We compare the correlation of the utilities for Sparse and Express BWS treatment cells using a test for independent correlations (Cohen and Cohen, 1983).

3.2. Results

3.2.1. Comparison of time statistics per experiment

The three BWS tasks took respondents about the same amount of time to answer. Median completion times for the nine Sparse and Express BWS questions were 2.63 and 2.57 min, respectively. For the full set of 27 questions, median completion times were 7.57 min for Full BWS, 7.28 min for Sparse BWS and 7.43 min for Express BWS. Though other questions came before the BWS section, fatigue should not have affected respondents because the median total survey length was just 7.9 min.

3.2.2. Comparison of Sparse and Express BWS to full BWS experiment

The first step is to compare the sample utilities found in the Sparse and Express BWS experiments with the sample utilities found in the control cell (the Full BWS experiment) using the test of dependent correlations (Cohen and Cohen, 1983).

While both Sparse and Express predict the Full BWS utilities and rank data well, the Sparse utilities may be more strongly correlated with the Full BWS utilities than are the Express utilities. This claim is significant at the 90% confidence interval in favor of Sparse BWS when comparing the difference between correlations for the utilities ($z = 1.93$, $p = 0.0\,536$). This confirms the Serpetti et al. (2016) findings that there are meaningful differences between Sparse and Express in terms of predicting utilities. However, the difference between correlations for ranks is not significant ($z = 1.27$, $p = 0.2041$) (Table 5).

3.2.3. Ability of Sparse and Express BWS to recover “known” utility estimates

In order to test the ability of Sparse and Express BWS to recover “known” parameters at the individual respondent-level, we compare the “known” utilities from the 27 Sparse and Express BWS quads, with the previously discovered utilities from the nine

Table 5
Sample level correlation with full BWS experiment.

Sample Results	Corr (utilities)	Corr (ranks)
Sparse BWS	0.9602	0.9416
Express BWS	0.9004	0.8937

Table 6
Respondent-level correlations with "known" parameter estimates.

	Sparse (27 quads)	
Individual Results Sparse (9 quads)	Corr (utilities) 0.8417	Corr (ranks) 0.8084
	Express (27 quads)	
Individual Results Express (9 quads)	Corr (utilities) 0.6712	Corr (ranks) 0.6026

Sparse BWS quads and the nine Express BWS quads.

Comparisons using an independent correlations test (Cohen and Cohen, 1983) find that the correlations between Sparse BWS utilities and the Sparse "known" utilities, as well as the Sparse BWS ranks and the Sparse "known" ranks, are significantly higher than the correlations between the Express BWS and Express "Known" utilities and ranks (Utility $z = 35.22$, $p < 0.001$; Rank $z = 36.16$, $p < 0.001$) (see Table 6).

This drastic difference reaffirms the advantage in favor of Sparse over Express BWS for large item sets. These comparisons also show that the decrement in quality is severe when choosing Express BWS over Full BWS when needing individual respondent-level utilities, but much less severe when choosing Sparse BWS over Full BWS.

4. Summary and conclusion

In this paper we test two methods for accommodating large numbers of items in a BWS experiment. Our motivation for this research comes from our role as practitioners, facing a common situation wherein our clients want to evaluate many items in their BWS studies. At the same time, those clients typically want to have respondent-level utilities capable of supporting subsequent analyses such as segmentation, simulations, and TURF analyses.

This research confirms consistent earlier findings about the equivalent ability of Sparse BWS and Express BWS to replicate sample level utilities (and sample item ranks) of a holdout set of respondents. The unique contribution is that this paper extends to human respondents the finding that Sparse BWS better reproduces known respondent-level utilities (and item ranks) than does Express BWS, a finding previously only shown for artificial respondents. Practitioners facing the common challenge of a BWS study with a large number of items can now make evidence-based decisions about how best to design their studies.

5. Discussion and next steps

5.1. When to use each method?

Some applied choice modelers prefer Sparse BWS, some prefer Express BWS and others are unsure which to use. Such preferences, when they exist, appear to be based more on hunches than on conceptual grounds. One assumption may be that Sparse BWS outperforms Express BWS because, although the data matrix is sparse at the individual level, the model has some information to inform each individual's utilities, where Express BWS has to rely more heavily on aggregate priors to estimate individual respondent-level utilities for the other items. On the other hand, some assume Express will outperform Sparse because more observations of each item at the individual level make the model's estimates more stable, at least for those items within the subset shown in the Express BWS experiment.

This research provides evidence-based methodological guidance to practitioners. Applied choice modelers can be reassured of the feasibility of commercial BWS studies with dozens or scores of items. Sparse and Express BWS both provide excellent sample level utility estimates. Moreover, this research provides evidence to practitioners that Sparse BWS produces much better individual respondent-level utilities than does Express BWS, and is thus a better choice if their study objectives involve estimating individual respondent-level item scores.

If a study's objectives require item scores only at the sample level and not for individual respondents, Full, Sparse, or Express BWS will fit the bill, with the latter two methods doing so with less effort on the part of the respondents and with less investment in questionnaire real estate. For research efforts that require individual respondent-level BWS utilities, Sparse BWS will outperform Express BWS in terms of producing more valid respondent-level utilities.

We do not believe that Sparse BWS scales up indefinitely. At some point beyond 100 items, when even Sparse BWS would require more than 30 or 40 choice sets, Sparse BWS may become too onerous for respondents and Express BWS may be the last method standing.

5.2. Psychological or algorithmic effects?

While our recommendations in favor of Sparse BWS are strong, we are unclear whether the advantages of Sparse result from psychological effects or algorithmic effects. For example, respondents are shown a greater variety of items in the Sparse approach,

versus a restricted subset of items in the Express approach. Therefore Sparse BWS may retain a respondent's interest better, resulting in better predictions. Algorithmically, Sparse BWS gives the Bayesian MNL model some information on every item, while Express offers three times as much information about a subset of the items, but no information on the remaining items. Our research does not identify the reason for the superior performance of Sparse BWS. Future research could investigate why Sparse performs better than Express BWS.

5.3. Other ways to improve express BWS?

Additional consideration should be given to using covariates within an Express BWS model, as borrowing information in a hierarchical model from other like-minded individuals may better inform the model for those items not shown in the subset. However, careful thought must be given while developing the questionnaire to ensure that there are proper variables in the study that would allow for this investigation. This research did not anticipate the need for covariates, and thus did not have useful options for investigation.

We are unsure as to which of Sparse BWS or Express BWS, if either, would favor different kinds of items, different mixes of more and less appealing items or even different dimensions on which the items might be scaled (e.g. preference, appeal, agreement, willingness to buy, etc.). Other considerations include enlarging the proportion of items included in the Express BWS. This study and Serpetti et al. findings only show that Sparse BWS outperforms Express BWS when each respondent sees a third of the total number of items under study in their Express BWS experiment.

5.4. Examining survey fatigue

While this paper was not designed to test survey fatigue and its impact on the performance of the Sparse and Express BWS approaches, we believe this is an interesting topic for future research and recommend an item set larger than 36, preferably closer to 100. Additional recommendations on different topics, scales, and size of the Express subset could also impact these measures.

5.5. BWS cases 2 and 3

The challenge posed by large numbers of items would face BWS Case 2 studies if profiles included many attributes. Likewise, for Case 3, the profile case, many attributes, and many profiles could both prove challenging. This paper does not specifically address these challenges for Cases 2 or 3.

Appendix

The results shown in the subsequent tables (Table 7 through 11) include three results – the sample (average) utility, the utility standard deviation, and the utility rank. The sample (average) utility is created by employing a hierarchical Bayesian (HB) model to estimate individual respondent-level utilities under the logit rule. Those scores, typically consist of both negative and positive values that are on an interval scale. Therefore, these scores are typically converted into probabilities that range from 0 to 100, with ratio-scaling properties where an item with a score of 4 is twice as preferred as an item with a score of 2. To convert the raw scores to the 0–100 point scale, one must first zero-center the individual scores by subtracting the mean score for each respondent from each respondent's scores. These zero-centered scores are then exponentiated per respondent using the following formula:

$$\frac{e^{U_i}}{(e^{U_i} + a - 1)}$$

where:

U_i = zero-centered raw logit weight for item i
e^{Ui} is equivalent to taking the antilog of U_i. In Excel, use the formula = EXP(U_i)
a = Number of items shown per set

This results in a score for all 36 items whose total scores sums to 100. After rescaling, one can take the standard deviation across the sample (a reflection of heterogeneity). The standard deviation is naturally larger for higher-scoring items and smaller for lower-scoring items. Finally, a rank is provided, transforming the sample (average) utilities into an overall Utility Rank.

For convenience, all the tables are sorted from best to worst with regard to the sample average utility scores for the Full BWS experiment.

Table 7

Full BWS Experiment Results (Sorted by Utility Rank) (n = 403)

Item #	Item	Sample (Average) Utility	Utility Standard Deviation	Utility Rank
26	New York cheesecake	4.773	2.536	1
29	Hot brownie sundae	4.544	2.175	2
9	Chocolate molten cake	4.319	2.357	3
13	Hot fudge sundae	4.256	2.129	4
27	Oreo cheesecake	4.252	2.566	5
17	Chocolate layer cake	4.112	2.252	6
11	Chocolate mousse	3.594	2.176	7
33	Red velvet cake	3.464	2.552	8
20	Skillet chocolate chip cookie	3.309	2.333	9
8	Apple pie	3.224	2.335	10
28	Peanut butter cheesecake	3.150	2.761	11
14	Salted caramel sundae	3.068	2.137	12
24	Peach cobbler	3.007	2.545	13
3	Belgian waffle ice cream cone	2.916	2.008	14
23	Louisiana mud pie	2.895	2.174	15
32	Tiramisu	2.861	2.684	16
31	Key lime pie	2.733	2.619	17
18	Lemon meringue pie	2.730	2.627	18
1	Crème brûlée	2.618	2.521	19
22	Texas pecan pie	2.607	2.562	20
5	French silk pie	2.510	1.932	21
2	Italian gelato	2.504	2.227	22
35	Cannoli	2.489	2.402	23
7	Coconut creme pie	2.398	2.557	24
6	Cinnamon roll	2.397	2.156	25
25	Cinnamon crumble cake	2.312	1.865	26
36	Pumpkin cheesecake	2.264	2.390	27
21	Cherry pie	2.185	2.257	28
15	Vanilla milk shake	2.178	2.092	29
16	Strawberry milk shake	1.999	2.128	30
34	Blueberry tart	1.790	1.940	31
4	Churros	1.714	1.980	32
19	Flan	1.469	2.063	33
12	Bread pudding	1.305	1.868	34
10	Mango lassi	1.120	1.712	35
30	Rice pudding	0.935	1.478	36

Table 8
Sparse BWS Experiment Results (9 quads) (Sorted by Full BWS Utility Rank) (n = 400)

Item #	Item	Sample (Average) Utility	Utility Standard Deviation	Utility Rank
26	New York cheesecake	4.462	2.467	5
29	Hot brownie sundae	5.366	2.214	1
9	Chocolate molten cake	5.003	2.402	3
13	Hot fudge sundae	4.753	1.911	4
27	Oreo cheesecake	4.256	2.393	6
17	Chocolate layer cake	5.252	2.016	2
11	Chocolate mousse	4.233	2.001	7
33	Red velvet cake	3.345	1.993	10
20	Skillet chocolate chip cookie	3.684	1.759	8
8	Apple pie	3.480	2.120	9
28	Peanut butter cheesecake	3.128	2.752	13
14	Salted caramel sundae	3.239	2.067	12
24	Peach cobbler	2.843	2.188	15
3	Belgian waffle ice cream cone	3.310	1.870	11
23	Louisiana mud pie	2.855	1.794	14
32	Tiramisu	2.202	1.747	26
31	Key lime pie	2.521	2.470	18
18	Lemon meringue pie	2.163	2.180	27
1	Crème brûlée	2.379	2.214	21
22	Texas pecan pie	2.221	1.991	25
5	French silk pie	2.712	1.720	16
2	Italian gelato	2.437	1.574	20
35	Cannoli	2.474	1.835	19
7	Coconut creme pie	2.347	2.316	22

(*continued on next page*)

Table 8 (*continued*)

Item #	Item	Sample (Average) Utility	Utility Standard Deviation	Utility Rank
6	Cinnamon roll	2.336	2.070	23
25	Cinnamon crumble cake	2.581	1.819	17
36	Pumpkin cheesecake	1.626	1.698	31
21	Cherry pie	1.917	1.736	28
15	Vanilla milk shake	2.315	1.793	24
16	Strawberry milk shake	1.673	1.510	29
34	Blueberry tart	1.511	1.685	32
4	Churros	1.663	1.693	30
19	Flan	0.924	1.377	34
12	Bread pudding	1.301	1.423	33
10	Mango lassi	0.684	0.932	36
30	Rice pudding	0.806	1.390	35

Table 9
Express BWS Experiment Results (9 Quads) (Sorted by Full BWS Utility Rank) (n = 404)

Item #	Item	Sample (Average) Utility	Utility Standard Deviation	Utility Rank
26	New York cheesecake	5.043	2.271	4
29	Hot brownie sundae	5.961	1.935	1
9	Chocolate molten cake	5.701	2.025	2
13	Hot fudge sundae	4.421	1.817	8
27	Oreo cheesecake	4.961	2.301	5
17	Chocolate layer cake	4.850	2.050	6
11	Chocolate mousse	5.662	1.510	3
33	Red velvet cake	3.201	2.003	14
20	Skillet chocolate chip cookie	4.521	2.256	7
8	Apple pie	3.327	2.085	10
28	Peanut butter cheesecake	3.033	2.613	16
14	Salted caramel sundae	3.290	2.073	11
24	Peach cobbler	2.149	2.000	19
3	Belgian waffle ice cream cone	3.286	2.006	12
23	Louisiana mud pie	3.409	2.177	9
32	Tiramisu	3.075	2.418	15
31	Key lime pie	1.853	2.476	23
18	Lemon meringue pie	1.345	1.769	31
1	Crème brûlée	1.814	1.915	25
22	Texas pecan pie	1.824	1.840	24
5	French silk pie	2.941	1.932	17
2	Italian gelato	2.206	1.992	18
35	Cannoli	2.116	1.846	20
7	Coconut creme pie	1.705	2.034	27
6	Cinnamon roll	2.062	2.050	21
25	Cinnamon crumble cake	3.285	2.146	13
36	Pumpkin cheesecake	1.170	1.619	33
21	Cherry pie	1.488	1.875	29
15	Vanilla milk shake	2.013	1.708	22
16	Strawberry milk shake	1.751	1.604	26
34	Blueberry tart	1.279	1.541	32
4	Churros	1.422	1.682	30
19	Flan	0.772	1.256	35
12	Bread pudding	1.532	2.014	28
10	Mango lassi	0.782	1.063	34
30	Rice pudding	0.751	1.284	36

Table 10
Sparse "Known" BWS Experiment Results (27 quads) (Sorted by Full BWS Utility Rank) (n = 400)

Item #	Item	Sample (Average) Utility	Utility Standard Deviation	Utility Rank
26	New York cheesecake	4.449	2.538	5
29	Hot brownie sundae	4.921	2.049	1
9	Chocolate molten cake	4.724	2.263	2
13	Hot fudge sundae	4.635	2.123	3

(*continued on next page*)

Table 10 (*continued*)

Item #	Item	Sample (Average) Utility	Utility Standard Deviation	Utility Rank
27	Oreo cheesecake	4.142	2.467	6
17	Chocolate layer cake	4.609	2.113	4
11	Chocolate mousse	3.949	2.143	7
33	Red velvet cake	3.246	2.402	11
20	Skillet chocolate chip cookie	3.436	2.219	8
8	Apple pie	3.337	2.361	9
28	Peanut butter cheesecake	3.031	2.673	13
14	Salted caramel sundae	3.250	2.389	10
24	Peach cobbler	2.911	2.663	14
3	Belgian waffle ice cream cone	3.203	2.045	12
23	Louisiana mud pie	2.630	2.027	18
32	Tiramisu	2.504	2.622	21
31	Key lime pie	2.636	2.637	17
18	Lemon meringue pie	2.324	2.399	26
1	Crème brûlée	2.391	2.418	24
22	Texas pecan pie	2.409	2.554	23
5	French silk pie	2.709	2.115	15
2	Italian gelato	2.495	2.008	22
35	Cannoli	2.549	2.203	19
7	Coconut creme pie	2.349	2.452	25
6	Cinnamon roll	2.527	2.211	20
25	Cinnamon crumble cake	2.692	2.036	16
36	Pumpkin cheesecake	2.006	2.384	29
21	Cherry pie	2.063	2.222	28
15	Vanilla milk shake	2.217	1.941	27
16	Strawberry milk shake	1.855	1.903	30
34	Blueberry tart	1.591	1.900	32
4	Churros	1.691	1.973	31
19	Flan	1.087	1.781	34
12	Bread pudding	1.446	2.020	33
10	Mango lassi	0.906	1.463	36
30	Rice pudding	1.079	1.817	35

Table 11
Express “Known” BWS Experiment Results (27 quads) (Sorted by Full BWS Utility Rank) (n = 404)

Item #	Item	Sample (Average) Utility	Utility Standard Deviation	Utility Rank
26	New York cheesecake	4.790	2.398	2
29	Hot brownie sundae	4.997	2.200	1
9	Chocolate molten cake	4.707	2.298	3
13	Hot fudge sundae	4.409	2.202	5
27	Oreo cheesecake	4.565	2.377	4
17	Chocolate layer cake	4.400	2.193	6
11	Chocolate mousse	4.086	2.088	7
33	Red velvet cake	3.497	2.316	9
20	Skillet chocolate chip cookie	3.825	2.334	8
8	Apple pie	2.953	2.296	12
28	Peanut butter cheesecake	2.858	2.670	15
14	Salted caramel sundae	3.365	2.405	10
24	Peach cobbler	2.789	2.573	18
3	Belgian waffle ice cream cone	2.839	2.031	16
23	Louisiana mud pie	2.934	2.170	13
32	Tiramisu	3.031	2.752	11
31	Key lime pie	2.158	2.480	26
18	Lemon meringue pie	2.020	2.298	28
1	Crème brûlée	2.481	2.321	19
22	Texas pecan pie	2.275	2.316	24
5	French silk pie	2.810	2.042	17
2	Italian gelato	2.294	2.139	23
35	Cannoli	2.372	2.049	21
7	Coconut creme pie	2.177	2.478	25
6	Cinnamon roll	2.476	2.115	20
25	Cinnamon crumble cake	2.900	2.181	14
36	Pumpkin cheesecake	1.927	2.230	29
21	Cherry pie	1.721	2.059	30

(*continued on next page*)

Table 11 (*continued*)

Item #	Item	Sample (Average) Utility	Utility Standard Deviation	Utility Rank
15	Vanilla milk shake	2.294	2.151	22
16	Strawberry milk shake	2.115	2.053	27
34	Blueberry tart	1.556	1.773	33
4	Churros	1.610	1.780	32
19	Flan	1.157	1.772	34
12	Bread pudding	1.640	2.143	31
10	Mango lassi	0.872	1.327	36
30	Rice pudding	1.103	1.809	35

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jocm.2019.01.002.

References

Allenby, G.M., Ginter, J.L., 1995. Using extremes to design products and segment markets. J. Market. Res. 32 (November), 392–403.

Bacon, L., Lenk, P., Seryakova, K., Veccia, E., 2007. Making MaxDiff more informative: statistical data fusion by way of latent variable modeling. In: 2007 Sawtooth Software Conference Proceedings. Sawtooth Software, Orem, pp. 327–343.

Baumgartner, H., Steenkamp, J.B., 2001. Response styles in marketing research: a cross-national investigation. J. Market. Res. 38, 143–156.

Chrzan, K., 2010. Using partial profile choice experiments to handle large numbers of attributes. Int. J. Market. Res. 52, 827–840.

Chrzan, K., 2015. A parameter recovery experiment for two methods of MaxDiff with many items. Sawtooth Software Research Paper available at: https://www.sawtoothsoftware.com/support/technical-papers/175-support/technical-papers/maxdiff-best-worst-scaling/1493-a-parameter-recovery-experiment-for-two-methods-of-maxdiff-with-many-items.

Cohen, J., Cohen, C., 1983. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 2d Ed. Lawrence Erlbaum, Hillsdale.

DeSarbo, W.S., Ramanswamy, V., Cohen, S.H., 1995. Market segmentation with choice-based conjoint analysis. Market. Lett. 6, 137–148.

Dyachenko, T., Reczek, R.W., Allenby, G.M., 2014. Models of sequential evaluation in best-worst choice tasks. Market. Sci. 33, 828–848.

Finn, A., Louviere, J.J., 1992. Determining the appropriate response to evidence of public concern: the case of food safety. J. Publ. Pol. Market. 11 (1), 12–25.

Green, P.E., Goldberg, S.M., Montemayor, M., 1981. A hybrid utility estimation model for conjoint analysis. J. Market. 45, 33–41.

Kuhfeld, W.F., Wurst, J.C., 2012. An overview of the design of stated choice experiments. In: 2012 Sawtooth Software Conference Proceedings. Sawtooth Software, Orem, pp. 165–199.

Lipovetsky, S., Conklin, M.W., 2014. Best-Worst scaling in analytical closed form solution. Journal of Choice Modeling 10, 60–68.

Lee, J.A., Soutar, G., Louviere, J.J., 2007. Measuring values using best-worst scaling: the LOV example. Psychol. Market. 24 (12), 1043–1058.

Lockshin, L., Cohen, E., 2015. How consumers choose wine: using best-worst scaling across countries. In: Louviere, J.J., Flynn, T.N., Marley, A.A.J. (Eds.), *Best-Worst Scaling: Theory, Methods And Applications* (159-176). Cambridge University, Cambridge.

Louviere, J.J., Flynn, T.N., Marley, A.A.J., 2015. Best-Worst Scaling: Theory, Methods and Applications. Cambridge University, Cambridge.

Marley, A.A.J., Islam, T., Hawkins, G.E., 2016. A formal and empirical comparison of two score measures for Best-Worst Scaling. Journal of Choice Modeling 21, 15–24.

Marley, A.A.J., Louviere, J.J., 2005. Some probabilistic models of best, worst, and best–worst choices. J. Math. Psychol. 49 (6), 464–480.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

Orme, B., 2005. Accuracy of HB estimation in MaxDiff experiments. Sawtooth Software Research Paper accessed online at. https://sawtoothsoftware.com/support/technical-papers/maxdiff-Best-Worst-scaling/accuracy-of-hb-estimation-in-maxdiff-experiments-2005 on 5-24-2017.

Sawtooth Software, 2013. The MaxDiff System Technical Paper. Sawtooth Software Technical Paper accessed online at. https://www.sawtoothsoftware.com/download/techpap/maxdifftech.pdf on 4-10-2018.

Serpetti, M., Gilbert, C., Peitz, M., 2016. "The Researcher's paradox: a further look on the impact of large scale choice exercises. In: 2016 Sawtooth Software Conference Proceedings, pp. 147–162.

Thorndike, Edward L., 1920. A constant error in psychological ratings. J. Appl. Psychol. 4, 25–29.

Wirth, R., Wolfrath, A., 2012. Using MaxDiff for evaluating very large sets of items. In: 2012 Sawtooth Software Conference Proceedings, pp. 59–78.

Zhang, J., Reed Johnson, F., Mohamed, A.F., Hauber, A.B., 2015. Too many attributes: a test of the validity of combining discrete choice and Best-Worst scaling data. Journal of Choice Modeling 15, 1–13.