

USING SEEDED ITEMS TO IMPROVE EXPRESS BEST WORST DESIGNS

THOMAS EAGLE
EAGLE ANALYTICS OF CA
JON GODIN
MEGAN PEITZ
NUMERIOUS INC.

EXECUTIVE SUMMARY

Previous research on many-item Best Worst tasks has shown that Sparse designs (where each item is seen only 1x per respondent, but all items are shown) have generally outperformed Express designs (where each item is seen 3x, but only a subset of the items [30%–50%] are shown per respondent), especially regarding out-of-sample predictions. At the 2023 Turbo Choice Workshop, a suggestion was made to potentially improve Express BW designs by including a fixed number of items [3–5] across all respondents, with the remainder selected randomly via a blocked design.

We tested this approach via both simulated data tests and a comparative exercise among live respondents. For the latter, we compared in-sample and out-of-sample predictions across five design cells: traditional BestWorst, Sparse BestWorst, traditional Express BestWorst, and two Express BestWorst cells using user-selected seeded items or informed seeded items. Unfortunately, despite our best hopes, we saw no improvement of the seeded item Express designs over the traditional Express designs, all of which were generally outperformed by the traditional BestWorst and Sparse BestWorst approaches. The lone exception came from the simulated data, where we observed that using seeded items can help when there is a lot of response error in the data.

BACKGROUND AND MOTIVATION

There are several methods for handling many items in MaxDiff analyses. For example, there are Sparse MaxDiff, Express MaxDiff, and the Thompson Sampling¹ approaches to handling many items. Sparse MaxDiff and Express MaxDiff are both variations of the traditional MaxDiff (Maximum Difference or Best-Worst) scaling method used in market research to determine the relative importance or preference of multiple items.

Sparse MaxDiff focuses on reducing the number of comparisons each respondent must make. Instead of evaluating all possible pairs of items, respondents are shown a series of tasks where every item is seen a minimum of 1 time. Many versions of the tasks are shown across respondents so that every pair of items is seen across the entire sample, but not within each respondent. This reduces the cognitive load on participants while still providing robust data on preferences and importance. By presenting fewer comparisons, Sparse MaxDiff aims to maintain data quality and reliability, even with fewer data points, making it efficient and suitable for

¹ Sawtooth Software offers a Thompson sampling method of MaxDiff they call Bandit Best-Worst. It uses Thompson Sampling to oversample the best items from previous respondents. We do not cover this approach in this paper.

surveys with many items or when respondent fatigue is a concern. The main drawback of the Sparse MaxDiff approach is that each item is seen only once and not in every possible pairwise comparison.

Express MaxDiff, on the other hand, is designed to expedite the MaxDiff process by showing each respondent a subset of the total number of items. It can be argued that using a subset of items is easier for the respondent to handle cognitively. Psychological literature suggests respondents have a difficult time evaluating too many attributes (for example, the recommendation to show respondents only 6-12 attributes in a conjoint choice design). By using subsets of items, Express MaxDiff allows each item in the subset to be seen multiple times and in more pairwise combinations. It also reduces the cognitive load on the respondent compared to Sparse MaxDiff because fewer items are seen by each respondent. Across the sample, many versions of the subsets of items are used such that every item, and more pairwise comparisons, are seen more times.

Both approaches utilize the power of hierarchical Bayes multinomial logit modeling to impute the parameters associated with the items, both at the individual respondent level and the aggregate, upper level, model.

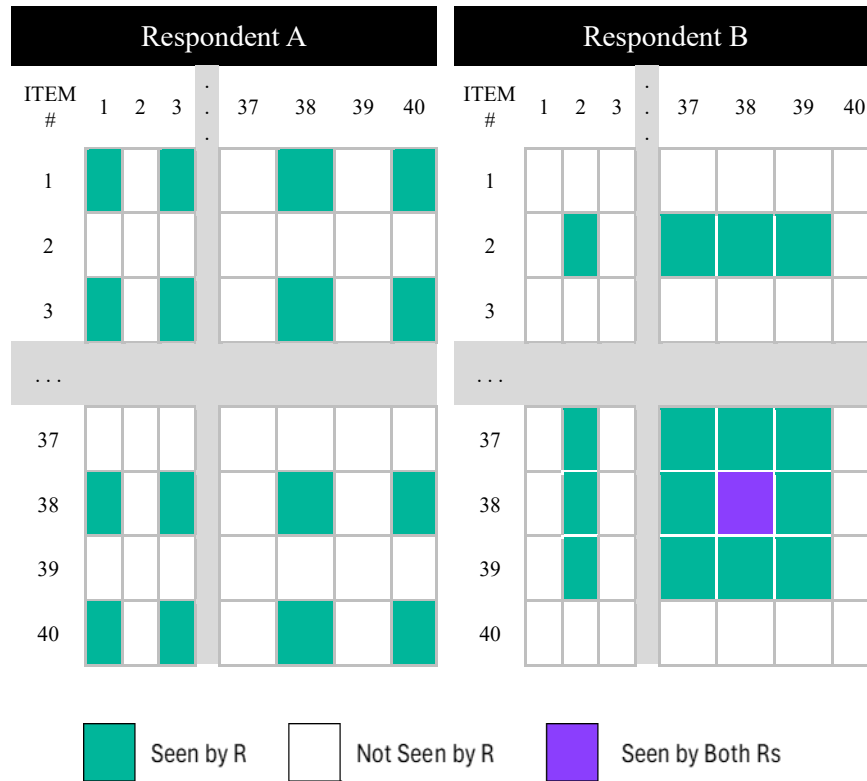
In comparison, while both methods aim to enhance the efficiency of the MaxDiff process, Sparse MaxDiff primarily reduces the number of item comparisons to lessen respondent burden, whereas Express MaxDiff focuses on collecting more data per item per respondent. Both approaches seek to balance the trade-off between data quality and respondent effort, albeit through slightly different mechanisms. In terms of predictive accuracy, Sparse MaxDiff has been shown to predict better to out-of-sample holdout tasks than Express MaxDiff

The idea of using a set of seeded items in an Express MaxDiff project arose during discussions at the 2023 Turbo Choice Modeling workshop where Tom Eagle discussed using them in some projects he had completed. Bryan Orme suggested a detailed evaluation of the use of seeded items in Express MaxDiff studies be conducted and presented at Sawtooth Software's A&I conference.

There is a lot of literature in psychology about the respondent's inability to cognitively process more than 15 to 20 items. Yet clients are very skeptical of using only a subset of items for each respondent; but showing many items only one time to each respondent also somehow feels inadequate. We will explore alleviating these concerns by including a small subset of items that will be seen by all respondents in an Express MaxDiff design.

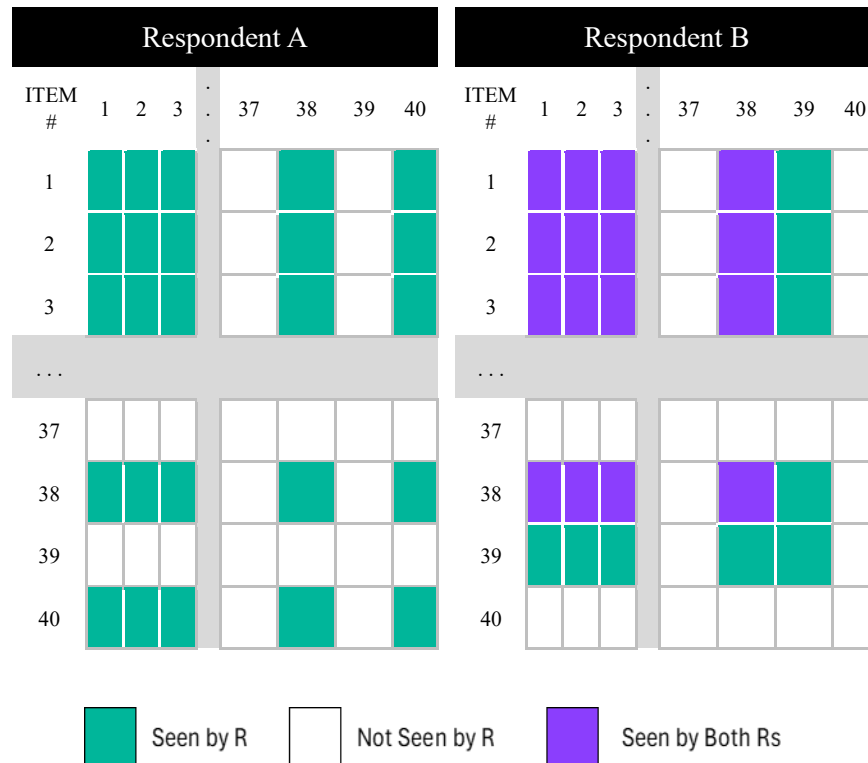
MOTIVATION FOR USING SEEDED ITEMS

The graphics below attempt to summarize the value of using seeded items. The tables below show what a traditional Express MaxDiff might show respondents.



The charts above show the pairs of items that might be seen in a traditional Express MaxDiff design. For two respondents: A sees items 1, 3, 38 and 40; B sees items 2 and 37 through 39. Only a single pair are seen by both respondents shown in purple.

In a seeded design, assuming 3 items are used as seeds, there is a minimum of 3 pairs of items seen by every respondent, and, as a result, more pairs seen in common across respondents A and B. The chart below shows the pairs seen in common are across items 1 to 3 and item 38



The upper-level portion of the HB MNL model of a seeded Express MaxDiff design has more data across pairs for use in estimating the variance-covariance matrix

SIMULATED RESPONDENT TESTING

We took an old 43 item Express MaxDiff study with 2,303 respondents as gospel. Their parameters are truth. We examine the traditional and seeded Express MaxDiff designs, as well as a Sparse MaxDiff design. All of these designs were generated using the recommendations of Wirth and Wolfrath (2012).

For the traditional Express MaxDiff we generated a design using 100 subsets of 20 items drawn from the 43 items. We generated 100 versions of 12 tasks with four items per task to show the 20 items.

In the seeded Express MaxDiff we generated 100 versions of 15 items, adding items 39–43 as the seeded items seen by every respondent. Again, we used 100 versions of 12 tasks with 4 items per task to show the 15+5 items.

The Sparse MaxDiff design had 100 versions of all 43 items. Each version had 9 tasks of 5 items in each task.

In all cases, respondents were randomly assigned to a single version in each task. We also tested each design using generated discrete choice responses with 3 values of the scale error (1,2, and 3).

The table below summarizes the design tested.

Traditional Design	# items/R	20
	# tasks	12
	# alts/task	5
	Scale	1, 2, 3
Seeded Design	# items/R	20
	# tasks	12
	# alts/task	5
	Seeded items	39–43
	Scale	1, 2, 3
Sparse Design	# items/R	43
	# tasks	9
	# alts/task	5
	Seeded items	NA
	Scale*	1, 2, 3
Runs	Burn-in	10,000
	Saved/10	1,000
	Prior df	50
	Prior Var	1.3

We estimated the hierarchical Bayes MNL model for every design using Sawtooth Software’s stand-alone CBC HB software to generate parameters. We used a burn-in of 10k iterations and we saved 1k iterations, saving every 10th iteration (10k total iterations after burn-in). Convergence was achieved quickly in every test.

For each Express MaxDiff design we predicted each respondent’s utility of every item in every task across every version. That is, each respondent was used to predict the utility of every item across the entire design, not just the version they were randomly assigned. This resulted in 13,818,000 predictions of utility and share (2,303 Rs x 100 versions x 12 tasks/version x 5 items/task).

In the Sparse MaxDiff design we also estimated the utility of every item across the entire design. This resulted in 10,363,500 predicted utilities and shares (2,303 Rs x 100 versions x 9 tasks/version x 5 items/task).

We compared the aggregated predicted to known parameters using correlations, mean absolute error (MAE) and compared the rankings of the top 5 and bottom 5 items. At the level of the individual respondents, we also compared predicted parameters and shares to the known respondent level parameters and shares. At the respondent level we examine the MAEs of each parameter (predicted to known), the MAEs of each parameter, the MAEs of the shares, and the MAE of the ranking of the items in each task. At our reviewer’s (Keith Chrzan) suggestion, we also examined the correlations of the predicted and known mean parameters across saved draws and across respondents for every parameter estimated.

SIMULATION RESULTS

The 2 tables below summarize the aggregate estimated to known parameter comparisons.

Correlations of the 43 Items			
	Scale = 1	Scale = 2	Scale = 3
Orig w/ Traditional	0.998	0.996	0.993
Orig w/ Seeded	0.998	0.997	0.993
Sparse	0.999	0.997	0.993

Mean Absolute Errors of the 43 Parameters			
	Scale = 1	Scale = 2	Scale = 3
Orig w/ Traditional	0.146	0.234	0.359
Orig w/ Seeded	0.162	0.238	0.328
Sparse	0.281	0.238	0.286

Looking at the correlation between the predicted and known parameters in the 1st table above we see very little difference among the three types of designs and across different scale error assigned to the generated choices.

The 2nd table of MAEs does show the traditional Express MaxDiff better predicts the actual parameters than the seeded Express and Sparse MaxDiff design. But the difference is relatively small. Interestingly, the Sparse MaxDiff had the largest MAEs when the scale was equal to 1 but performed as well or better than the other design as scale error increased.

The next set of tables compares the predicted versus the known rankings of the top and bottom 5 items

Top 5 Parameter Rankings			
	Scale = 1	Scale = 2	Scale = 3
Orig w/ Traditional	all 5	all 5	3 of 5
Perfect match	1	5	2
Orig w/ Seeded	4 of 5	4 of 5	all 5
Perfect match	1	0	3
Sparse	all 5	4 of 5	4 of 5
Perfect match	5	0	2

Bottom 5 Parameter Rankings

	Scale = 1	Scale = 2	Scale = 3
Orig w/ Traditional	4 of 5	all 5	4 of 5
Perfect match	4	3	4
Orig w/ Seeded	all 5	all 5	4 of 5
Perfect match	5	5	4
Sparse	4 of 5	all 5	4 of 5
Perfect match	4	5	2

These data suggest the traditional design is marginally better than the seeded design and about the same as the Sparse MaxDiff design when the scale error is 1. As scale error increases the predicted rankings remained about the same. A perfect match represents when the number of top and bottom 5 items were matched EXACTLY with the known parameters.

When examining the predicted parameters and shares across all the tasks in the entire designs, we do not see major differences across the three types of designs.

Average Aggregated Parameter MAE across all tasks

	Scale = 1	Scale = 2	Scale = 3
Orig w/ Traditional	1.008	0.777	0.725
Orig w/ Seeded	1.008	0.777	0.728
Sparse	1.101	0.838	0.747

The average aggregated parameter MAEs across all tasks across all items are about the same, with the Sparse MaxDiff design slightly worse. As scale increases, the parameters regress towards zero which results in lower MAEs as scale error increases. The same conclusions can be drawn in the average aggregated shares across all tasks in the table below.

Average Aggregated Share MAE across all tasks

	Scale = 1	Scale = 2	Scale = 3
Orig w/ Traditional	0.119	0.099	0.092
Orig w/ Seeded	0.116	0.097	0.092
Sparse	0.122	0.101	0.091

Examining the predicted to known ranking of items in each task across the entire design clearly shows the Sparse MaxDiff design performing better than either the traditional or seeded Express MaxDiff designs. The impact of increasing scale error is inconclusive. These MAEs are the predicted rank of the items in each task to the known ranking.

Lastly, we examine the average correlation of each of the predicted and known mean posterior draws.

Average Correlation Across 43 Items Mean Posterior Draws			
	Scale = 1	Scale = 2	Scale = 3
Orig w/ Traditional	0.687	0.499	0.361
Orig w/ Seeded	0.681	0.674	0.569
Sparse	0.675	0.456	0.325

At a scale error of 1 the best mean correlation across the 43 known and prediction parameters is the traditional Express MaxDiff design. The seeded Express MaxDiff and Sparse MaxDiff design are slightly lower. But, as scale increases the traditional and Sparse MaxDiff design correlations drop quickly and by a large amount, whereas the seeded Express MaxDiff design more closely predicts the known parameters. This suggests the seeded Express MaxDiff design can more closely reproduce the respondent heterogeneity in the posterior individual respondent level mean parameters. This is the only result that suggests the value of using a seeded Express MaxDiff design.

LIVE RESPONDENT TESTING

Research Plan

For our live respondent testing, we studied candy preferences among a set of 40 different candies:



Traditional approaches to testing preferences for large item sets like this include **standard BestWorst designs** where each item is shown ~3 times to each respondent, but all items are shown; **Sparse BestWorst designs** where each item is shown only 1 time to each respondent, but all items are shown; and **Express BestWorst designs** where each item is shown ~3 times to each respondent, but only a randomly-selected subset of items (typically one-third to one-half of the items) are shown to each respondent. (See Chrzan and Peitz, 2019; Godin et al., 2023, Orme, 2019; Serpetti et al., 2016; and Wirth and Wolfrath, 2012).

In addition to using each of those approaches as a test cell, we created two experimental cells based on an Express BestWorst framework but employing the use of a small subset of fixed item “seeds” that would be seen by all respondents in the cell, with the remaining items selected at random. Our first experimental approach tests the use of “user selected” seeds, designed to mimic a situation where a client selects the items they want all respondents to see—we had no client, so we just used the first five candies as seeds for this cell. Our second experimental approach tests whether using “informed seeds” that span the top, middle, and bottom of the preference spectrum perform better than the user selected seeds. Here, we collected a pretest census-balanced sample of 50 respondents using a traditional Express BestWorst design and estimated their utilities using a hierarchical Bayes model; for subsequent respondents, we used the candies with mean probability scores ranked #1, #2, # 20, #39, and #40 from the pretest sample HB results as the fixed seeds. Cells 3-5 all show only 20 of the 40 items to each respondent.

#	Approach	Task Structure	Size	Description
1	Traditional BestWorst Exercise	- 5 items per task - All 40 items included - Each item shown 3x per respondent	# Tasks: 24 N Size = 301	Standard full design
2	Traditional Sparse BestWorst Exercise	- 5 items per task - All 40 items included - Each item shown 1x per respondent	# Tasks: 8 N Size = 300	Standard Sparse design
3	Traditional Express BestWorst Exercise	- 5 items per task - 20 items randomly selected per respondent - Each item shown 3x per respondent	# Tasks: 12 N Size = 316	Standard Express design with no seeded items
4	Express BestWorst with User-Selected Seeds	- 5 items per task - 20 items per respondent: 5 fixed, 15 randomized - Each item shown 3x per respondent	# Tasks: 12 N Size = 309	Designed to mimic a client selecting certain items to use as seeds. We just used the first 5 items as seeds here.
5	Express BestWorst with Informed Seeds	- 5 items per task - 20 items per respondent: 5 fixed, 15 randomized - Each item shown 3x per respondent	# Tasks: 12 N Size = 308	HB utilities were estimated from 50 census-balanced initial respondents; seeds for subsequent respondents were 1 st , 2 nd , 20 th , 39 th , and 40 th ranked items.

To validate our models, we first set up two BestWorst-style holdout tasks unique to each cell where each respondent within each cell saw the same two tasks. These holdout tasks were determined randomly by the Lighthouse Studio designer using one-version, two-task designs, each displaying five alternatives. Respondents were asked to select which candy they would be most likely to choose, and which candy they would be least likely to choose from the set shown. These fixed tasks are held out from estimation, and we used the utilities estimated from the main exercise for the cell to predict the holdout choices. An example task from Cell 1 (Traditional BestWorst) is shown below (note that the main tasks for all cells used this same task structure—all that varied across the cells was the number of items in the underlying design, the number of times each item was shown in that design, and the number of tasks shown to each respondent):

Imagine you are given the choice of eating any of the candies shown. Which would you say you'd be **MOST LIKELY** to choose, and which would you be **LEAST LIKELY** to choose?

(1 of 26)



Most Likely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Least Likely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>






Click the 'Next' button to continue...

Our second holdout approach utilized universal fixed ranking-based holdout tasks, where all respondents across all cells saw the same two tasks and were asked to rank each of the five candies shown in each task in order of preference, where 1=most preferred and 5=least preferred. Once again, we will use each set of estimated utilities for each cell to try to match the observed preferences from the holdout tasks, both within-sample (testing each cell's utility predictions of the holdout task choices for that cell) and out-of-sample (testing each cell's utility predictions of the holdout choices made by all the other cells). The two ranking holdout tasks used are shown below.

Next, here are five candies that you may or may not have seen previously. Please rank the candies in order of your personal preference, where 1 is the best and 5 is the worst.

Click on the item and drag it to the position of your choice.

Items to Rank

-  Kit Kat
-  Snickers
-  Jolly Ranchers
-  Lifesavers Big Ring Gummies
-  Baby Ruth






Best

Worst

Here are five other candies that you may or may not have seen previously. Please rank the candies in order of your personal preference, where 1 is the best and 5 is the worst.

Click on the item and drag it to the position of your choice.

Items to Rank

-  Swedish Fish
-  Sour Patch Kids
-  Hershey's Milk Chocolate
-  Reese's Peanut Butter Cup
-  Twizzlers

Best

Worst

FIELDWORK AND DATA CLEANING

We executed the survey in January and February 2024 using Prodege’s peeq marketplace sample, among respondents 18+ who eat candy at least a few times per year.

The data was cleaned for speeding (defined as those completing the survey in $< \frac{1}{2}$ the median completion time) and age mismatch between open-ended age asked early in the questionnaire and year of birth asked at the end of the questionnaire (respondents with a discrepancy of 2 or more years were removed).

We did not utilize any on-the-fly or post hoc Root Likelihood (RLH) based approaches to catching respondents who answer as randomly as true random robots (Chrzan and Orme, 2022), because we are utilizing sparse designs which produce less stable RLH estimates (typically inflated) since items are not seen more than once, so this type of data scrubbing would be imbalanced across design cells.

In the questionnaire we collected other data including candy purchase frequency, spending, attitudes towards sour candy and candy with nuts, whether the respondent has a nut allergy, and area, region, and state.

MODEL ESTIMATION

For each cell, we estimated the utilities twice using hierarchical Bayes (HB): first with no covariates included, and then with six covariates including age group, greater than median preference for chocolate candy, greater than median preference for hard candy, or greater than median preference for chewy candy, attitude towards sour candy, and attitude towards peanuts and nuts in candy.

For the MaxDiff estimation we used the standard Lighthouse Studio defaults of 20,000 burn-in iterations, and 10,000 posterior saved iterations.

ANALYSIS OF RESULTS

Candy Preferences

Our first point of comparison across the five design cells is to review the candy preferences themselves to see whether we get very different preference structures emerging from the varying designs.

In the table below are the estimated utilities from each model, transformed using the probability scaling approach where preferences across the items sum to 100, the scores are ratio scaled, and an item with a score of 4 is two times as preferable as an item with a score of 2. Scores have been sorted by the mean preferences of cell 1 (traditional BestWorst design), and results shown are from the models without covariates:

Top 20 Items	Item #	Cell 1: Traditional BestWorst	Cell 2: Traditional Sparse BW	Cell 3: Traditional Express BW	Cell 4: User Selected Seeded Express BW	Cell 5: Informed Seeded Express BW
Reese's Peanut Butter Cup	24	5.14	5.79	5.66	5.98	5.77
Kit Kat	14	4.79	4.79	4.77	5.00	5.32
Snickers	30	4.62	4.90	5.27	5.79	5.27
M&Ms	18	4.38	4.61	4.69	4.33	5.04
Hershey's Milk Chocolate	10	4.34	4.55	4.93	5.22	5.26
Twix	35	4.30	4.46	4.50	4.99	5.17
Milky Way	20	3.85	3.83	4.28	4.51	4.21
Reese's Pieces	25	3.82	3.39	3.42	4.33	4.03
3 Musketeers	1	3.72	3.32	3.41	3.64	3.98
Butterfinger	5	3.65	3.60	3.80	3.69	3.80
Lindt Dark Chocolate Truffles	17	3.35	3.63	3.95	4.00	3.59
Baby Ruth	4	2.92	2.97	3.30	3.22	2.98
Almond Joy	3	2.90	3.53	2.86	2.96	2.91
Rolo	26	2.71	2.91	3.10	2.86	3.08
York Peppermint Patties	40	2.66	2.34	2.50	2.38	3.04
Skittles	28	2.64	2.70	2.11	1.95	2.45
Heath Bar	9	2.63	3.03	2.77	3.19	2.58
Starburst	32	2.43	2.45	2.62	2.19	2.16
Payday	22	2.39	2.09	2.63	2.25	2.37
Whatchamacallit	38	2.32	2.58	2.65	2.73	2.79

Bottom 20 Items	Item #	Cell 1: Traditional BestWorst	Cell 2: Traditional Sparse BW	Cell 3: Traditional Express BW	Cell 4: User Selected Seeded Express BW	Cell 5: Informed Seeded Express BW
Lifesavers Big Ring Gummies	16	2.22	2.69	2.22	2.06	2.11
Haribo Gold Bears	8	2.18	2.34	1.76	1.75	1.62
Jolly Ranchers	12	2.13	1.79	1.61	1.15	1.40
Sour Patch Kids	31	2.13	2.06	2.42	1.85	1.96
Werther's Original	37	2.12	1.64	1.18	1.80	1.56
Junior Mints	13	1.91	1.42	2.22	1.95	1.32
Airheads	2	1.82	1.30	1.24	1.48	1.09
Milk Duds	19	1.74	1.76	2.08	1.57	2.05
Whoppers	39	1.66	1.57	1.42	1.67	1.98
Twizzlers	36	1.63	1.63	1.54	0.90	1.40
Jelly Belly 50 Flavor Jelly Beans	11	1.53	1.51	1.40	0.96	0.81
Skor	29	1.45	1.14	1.64	1.25	1.27
Swedish Fish	33	1.41	1.48	1.18	1.21	1.07
Lemonhead	15	1.26	1.30	0.93	0.90	0.94
Tootsie Rolls	34	1.16	1.28	1.04	1.37	0.99
Nerds	21	1.06	1.20	0.76	0.84	0.79
Dum Dums Lollipops	7	0.93	0.70	0.51	0.46	0.52
Dots	6	0.80	0.67	0.62	0.49	0.62
Pop Rocks	23	0.72	0.47	0.66	0.36	0.43
Runts	27	0.58	0.59	0.38	0.78	0.29

From a visual inspection of the heatmap, we observe that the darker orange (“hot”) colors are generally floating towards the top of the table across all cells, with the top six items being consistent (varying slightly in order only), with the exception of M&Ms for cell 4 being outside that cell’s top 6. Looking at the darker blue (“cold”) colors, we see that they also move consistently toward the bottom of the table, with the bottom 4 candies being universally the same, again with slight variations in preference order.

So, at least from a high-level visual perspective, each of the techniques is capturing roughly similar preferences toward the top and bottom of the scale, with some larger fluctuations in the middle. We can confirm that we are capturing similar preference structures a bit more quantitatively by comparing the correlations of the mean importance scores across the five cells, as shown in the table below (the highest correlation is indicated in **bold black**; the lowest correlation is shown in **bold red**):

Correlations between Importance Scores across design cells	Traditional BW	Traditional Sparse BW	Traditional Express BW	User Selected Seeded Express BW	Informed Seeded Express BW
Traditional BW	1.000	--	--	--	--
Traditional Sparse BW	0.979	1.000	--	--	--
Traditional Express BW	0.974	0.972	1.000	--	--
User Selected Seeded Express BW	0.973	0.968	0.976	1.000	--
Informed Seeded Express BW	0.975	0.972	0.972	0.975	1.000

Overall, the correlations are quite strong, with a low of 0.968 between cells 2 and 4, and a high of 0.979 between cells 1 and 2. At a very baseline level, none of the approaches generate drastically different preference data, which is perhaps most reassuring from a client’s perspective. But are we getting similar predictive fits to both in-sample or out-of-sample holdouts across the cells? We will examine those results next.

Aside on Model Validation

Before we get to those results, here is a little background in case you are not familiar with calculating Hit Rates and Mean Absolute Errors (MAEs) to see how well your data is predicting fixed holdout task choices.

For **Hit Rates**, we start by building a model to estimate individual-level utilities from our BestWorst exercises. Then, we simulate the holdout task options at the *individual* level and use the estimated utilities to predict which item the respondent will pick as the best and worst option of each holdout. We then compare the predicted choices against the observed holdout task choices: if the choices match, we count it as a hit, and if they do not match, we count it as a miss. Finally, we take the sum of the hits across all respondents divided by the total correct possible to get our overall Hit Rate values. Here is what these choices might look like for one respondent:

	Holdout 1		Holdout 2	
	Best	Worst	Best	Worst
Actual Choices	Item 1	Item 7	Item 12	Item 3
Simulated Choices	Item 1	Item 14	Item 12	Item 26
	✓	X	✓	X

For this respondent, we correctly predicted both of their Best choices, but we got both of the Worst choices wrong. So, the Hit Rate for this respondent is 50% (2 correct out of 4), where a perfect score would of course be 100%.

Next, here is how **Mean Absolute Error (MAE)** calculations work. Again, using our estimated BestWorst utilities, we simulate the holdout task now at the *aggregate* level. We then compare the share of preference predictions from the estimated utilities to the actual share of choice frequencies of each of the holdout tasks for the sample. We calculate the difference between the estimated and actual shares, and we take the absolute value of those differences and average across them—the result is the Mean Absolute Error.

	Holdout 1			
	Item 1	Item 2	Item 3	Item 4
Actual Choice Shares	20%	30%	10%	40%
Simulated Choice Shares	22%	28%	15%	35%
Differences	+2%	-2%	+5%	-5%
Absolute Differences	2	2	5	5
Mean Absolute Error	$(2 + 2 + 5 + 5)/4 = (14/4) = \mathbf{3.5}$			

In the example above, our share predictions are off by 2 points each for the first two items, and 5 points each for the second two items, resulting in an MAE of 3.5. Again, a perfect score would be zero, no error. As an analyst you can attempt to minimize the MAE by tuning the exponent (a constant value used to scale the utility values up [producing more differentiation in the data] via an exponent > 1, or scale them down [flattening the data] via an exponent < 1), but for this paper all MAEs reported are from their “natural,” un-tuned state.

In-Sample Holdout Validation

For the in-sample BestWorst holdouts, we start by examining the individual-level Hit Rates to see how well the model for each design cell captures individual preferences. We compare the observed best and worst choices from the holdout tasks of each respondent to the choices we predicted they would make based on their individual-level utilities. Note that the BestWorst-style holdout tasks were unique for each design cell, but all respondents within the cell saw the same two holdout tasks. We ran these comparisons twice, using each of the two sets of utilities for each cell: models estimated without covariates, and models estimated with covariates. A table of results is below.

In-Sample Hit Rate Comparison

Cell	NO COVARIATES					WITH COVARIATES				
	1 Traditional BW	2 Traditional Sparse BW	3 Traditional Express BW	4 User Selected Seeded Express BW	5 Informed Seeded Express BW	1 Traditional BW	2 Traditional Sparse BW	3 Traditional Express BW	4 User Selected Seeded Express BW	5 Informed Seeded Express BW
Holdout1 Best	66.8%	57.7%	57.0%	66.7%	58.1%	68.4%	55.0%	53.2%	59.5%	56.2%
Holdout 1 Worst	64.1%	62.0%	55.4%	64.7%	58.1%	63.8%	60.3%	57.3%	62.5%	59.4%
Holdout 2 Best	65.4%	57.3%	67.1%	51.8%	47.1%	64.5%	56.7%	63.9%	51.5%	50.6%
Holdout 2 Worst	69.1%	69.3%	43.7%	48.2%	56.8%	66.8%	66.7%	45.3%	51.1%	57.1%
Overall Best	66.1%	57.5%	62.0%	59.2%	52.6%	66.4%	55.8%	58.5%	55.5%	53.4%
Overall Worst	66.6%	65.7%	49.5%	56.5%	55.0%	65.3%	63.5%	51.3%	56.8%	58.3%
Overall	66.4%	61.6%	55.8%	57.8%	55.0%	65.9%	59.7%	54.9%	56.1%	55.8%
Difference with Covariates						-0.5%	-1.9%	-0.9%	-1.7%	+0.8%

Looking first at the results from the models with no covariates, the Traditional BestWorst design resulted in the highest hit rate, at 66.4% overall; the Traditional Sparse BestWorst design also does fairly well, but the hit rates for the three Express design cells are lower, with no observable improvement of either seeded approach over the Traditional Express model.

These results are consistent with expectations, and also reflect a slight methods bias in the data. For the Traditional BestWorst design, each respondent saw all 40 of the candies three times each, so we achieve more stable individual-level results using that approach. For the Traditional Sparse BestWorst design, items were shown only once, but all items were shown so there is no chance for accidental surprise items showing up in the holdout tasks. This is the bias facing the Express designs—since the items used for each respondent in the Express design cells are only a randomized subset of the total items in the design, there’s no guarantee that *any* of the items that were included in the holdout tasks were actually shown to a particular respondent. So, it’s not surprising that the scores for these cells are lower, but if our hypothesis were true, we would have expected that the Seeded Express designs would have higher scores than the traditional Express design, and the evidence does not support that conclusion.

When we look at the models that included covariates, we observe a similar story. The Traditional BestWorst design achieves the highest hit rates, followed by the Traditional Sparse BestWorst design, and then all of the Express cells do a little bit worse, with no meaningful improvement of the Seeded designs over the Traditional Express design. Consistent with previous research presented over the years at the Sawtooth Software Conference/Analytics & Insights Summit, the use of covariates did not improve the individual-level hit rates for the in-sample validation tests: four of the five cells actually display lower hit rates once covariates were included. Note that this does not mean that covariates are universally not helpful!—we will come back to this subject a bit later in the paper.

Continuing our examination of the in-sample BestWorst holdouts, we move next to the aggregate-level MAEs achieved by each of the models. Here we are trying to predict how the entire sample within each design cell will make choices in the holdout tasks for that cell, based on the utilities estimated for that cell. A table of results is provided below.

In-Sample MAE Comparison

Cell	NO COVARIATES					WITH COVARIATES				
	1 Traditional BW	2 Traditional Sparse BW	3 Traditional Express BW	4 User Selected Seeded Express BW	5 Informed Seeded Express BW	1 Traditional BW	2 Traditional Sparse BW	3 Traditional Express BW	4 User Selected Seeded Express BW	5 Informed Seeded Express BW
Holdout I MAE	2.7%	2.4%	1.8%	2.5%	3.6%	3.0%	3.3%	1.4%	2.7%	3.8%
Holdout I MAE	2.7%	2.8%	1.7%	4.2%	4.7%	2.8%	2.9%	1.3%	3.3%	4.9%
Overall	2.7%	2.6%	1.7%	3.4%	4.2%	2.9%	3.1%	1.4%	3.0%	4.3%
Difference with Covariates						+0.2%	+0.5%	-0.3%	-0.4%	+0.1%

Overall, the MAEs are relatively small for all models, so we are adequately capturing in-sample preferences with each model. As before, the Traditional BW and Traditional Sparse BW approaches tend to have the lowest error, especially in comparison to the two Seeded design cells. There is a bit of an unexpected anomaly for the cell 3 holdouts, as we have very low prediction error for that cell where we would have expected it to be more in-line with the other cells. We suspect that that may have been a lucky design where for whatever reason it was easier to predict those particular holdouts at the aggregate level than some of the other cells. Should anyone ever decide to repeat this study, we are highly dubious that they would see a similar result.

Focusing on the covariate side of the table, we again see a similar story. The traditional cells achieve a slightly lower error rate than the seeded Express cells, with cell 3 being an unexpected outlier. Most tellingly, however, we do not see either of the Seeded designs showing any improvement or ability to lower the error for an Express-type design.

As we saw with the in-sample Hit Rates, the in-sample MAEs generally do not improve much if at all when covariates are included in the model. Three of the five cells display slightly higher error rates with covariates than without.

Out-of-Sample (Ranking Task) Holdout Validation

For out-of-sample testing, we use a slightly different approach for our validation testing. Recall that for these holdouts we are not using standard BestWorst-style questions, but two ranking questions, each ranking five candies per screen in order of preference. Rather than trying to predict the accuracy of *all* the rankings for each holdout, which is a difficult hurdle to jump under the best of circumstances, we instead cycled through different iterations of groups of rankings, looking at each set of pairs, triples, quads and quints that emerge from the five items used in a given holdout ranking task.

In other words, for any given pair in the holdout, can we predict the relative first choice preference correctly? For any given set of three items can we predict the relative first choice correctly? For any given set of four items, can we predict the relative first choice correctly? And finally for the full set of five items, can we predict the relative first choice correctly?

For this data, we first examined the within-cell Hit Rates, using each cell’s utility data to predict its own observed holdout ranking task responses.

Ranking Task In-Sample Hit Rate Comparison

Cell	NO COVARIATES					WITH COVARIATES				
	1 Traditional BW	2 Traditional Sparse BW	3 Traditional Express BW	4 User Selected Seeded Express BW	5 Informed Seeded Express BW	1 Traditional BW	2 Traditional Sparse BW	3 Traditional Express BW	4 User Selected Seeded Express BW	5 Informed Seeded Express BW
Pairs (10 sets per holdout)	84.62%	79.52%	80.14%	78.19%	79.90%	84.30%	79.43%	78.83%	77.85%	78.59%
Triplets (10 sets per holdout)	78.95%	71.90%	71.85%	70.49%	73.10%	78.80%	71.82%	70.19%	69.37%	71.79%
Quads (5 sets per holdout)	74.29%	66.53%	65.51%	64.24%	68.77%	74.35%	66.57%	63.70%	61.91%	67.21%
Quints (1 set per holdout)	70.27%	62.50%	60.60%	58.09%	65.58%	70.43%	62.50%	58.70%	53.56%	64.45%
Weighted Average	79.90%	73.44%	73.39%	71.77%	74.59%	79.74%	73.38%	71.82%	70.59%	73.24%
Difference with Covariates						-0.16%	-0.06%	-1.56%	-1.18%	-1.35%

For the data with no covariates used in estimation, the Traditional BestWorst design achieves the highest hit rate, with the Traditional Sparse and Traditional Express designs producing similar results. While the user-selected seeded design fares the worst of all, in this case the informed seeded design does demonstrate slightly higher hit rates than the traditional express BestWorst design and even the Traditional Sparse design.

Moving to the results *with* covariates, we see the same pattern: Traditional BestWorst has the highest hit rates, the Sparse BestWorst cell is at par with the informed seeded model, and the user-selected seeded design performs the worst overall, if only by small margins. And once again we see that including covariates in the model does not lead to improved in-sample predictions, as all of the hit rates decrease slightly when covariates are used in the upper-level model during estimation.

Next we shift our focus to what we really care about the most—comparing how well our various models predict the rankings of respondents OUTSIDE of that particular cell. Out-of-sample validation is the gold standard for comparing model performance, as it is generally a higher-to-much-higher hurdle to cross.

The table below contains the MAEs for the different groupings of items within the holdout ranking questions for each of the design cells under study. For each column, we are using that cell’s utilities to predict the rankings made by respondents from all of the *other* cells combined.

Ranking Task Out-of-Sample MAE Comparison

Cell	NO COVARIATES					WITH COVARIATES				
	1 Traditional BW	2 Traditional Sparse BW	3 Traditional Express BW	4 User Selected Seeded Express BW	5 Informed Seeded Express BW	1 Traditional BW	2 Traditional Sparse BW	3 Traditional Express BW	4 User Selected Seeded Express BW	5 Informed Seeded Express BW
Pairs (10 sets per holdout)	2.9%	4.3%	5.0%	7.2%	5.5%	3.3%	4.0%	3.3%	5.5%	4.3%
Triplets (10 sets per holdout)	3.2%	4.7%	5.2%	7.2%	6.8%	3.7%	3.9%	3.2%	5.9%	5.4%
Quads (5 sets per holdout)	3.0%	4.8%	5.2%	7.0%	7.7%	3.3%	3.5%	3.0%	5.5%	5.7%
Quints (1 set per holdout)	2.5%	4.9%	5.3%	6.7%	8.6%	2.6%	3.3%	2.8%	5.2%	6.2%
Weighted Average	3.0%	4.6%	5.1%	7.1%	6.5%	3.4%	3.8%	3.2%	5.6%	5.1%
Difference with Covariates						+0.4%	-0.7%	-1.9%	-1.5%	-1.5%

Overall, the MAEs are relatively low. For the models without covariates, the Traditional BestWorst cell maintains the lowest out-of-sample error rate, followed by the Traditional Sparse BestWorst design. As has been shown in the studies cited previously in this paper, the Traditional Express design fares worse at out-of-sample prediction than Sparse designs do. And, once again, contrary to our hypothesis, including small subsets of fixed items (i.e., seeds) along with a randomly-drawn subset of items in an Express design does not appear to improve the results. This of course is only one study, but the out-of-sample error rates are substantially higher for the seeded designs, so using seeded items made things worse, not better.

Looking next to the results *with* covariates, a similar story emerges, but now the three Traditional cells are all more or less at par, while the seeded designs have quite a bit more error. As my kids would say, *whomp whomp*. Alas, our idea does not seem to improve out-of-sample performance of Express BestWorst designs.

However, it is exceedingly important to look at the relative performance of the models with covariates versus those without when it comes to validating out-of-sample choices. As my colleagues and I also showed in our 2023 paper (Godin et al., 2023), the covariates really do help lower the out-of-sample error for the more sparse design approaches used for understanding preferences with large sets of items, like Sparse BestWorst and Express BestWorst. While this current study shows relatively modest improvement compared to last year’s study (which dealt

with not only large item sets, but large item sets with very long statements [> 200 characters]), the improvement appears meaningful to us. The more difficult the task, and the more sparse you need to make the design, the more using (good) covariates in the model will help you better match out-of-sample preferences.

How to choose good covariates is a subject for another paper, but in this case we designed the questionnaire to only ask a small set of questions that should be highly-related to candy preferences: how strongly they favor chocolate vs. hard vs. chewy candy, whether they favor or are averse to sour candies or candies with nuts, and age, with different generational cohorts being drawn to different groups of candy, as tastes change over time (sour candies were quite rare when I was young, for example, and now you find them everywhere).

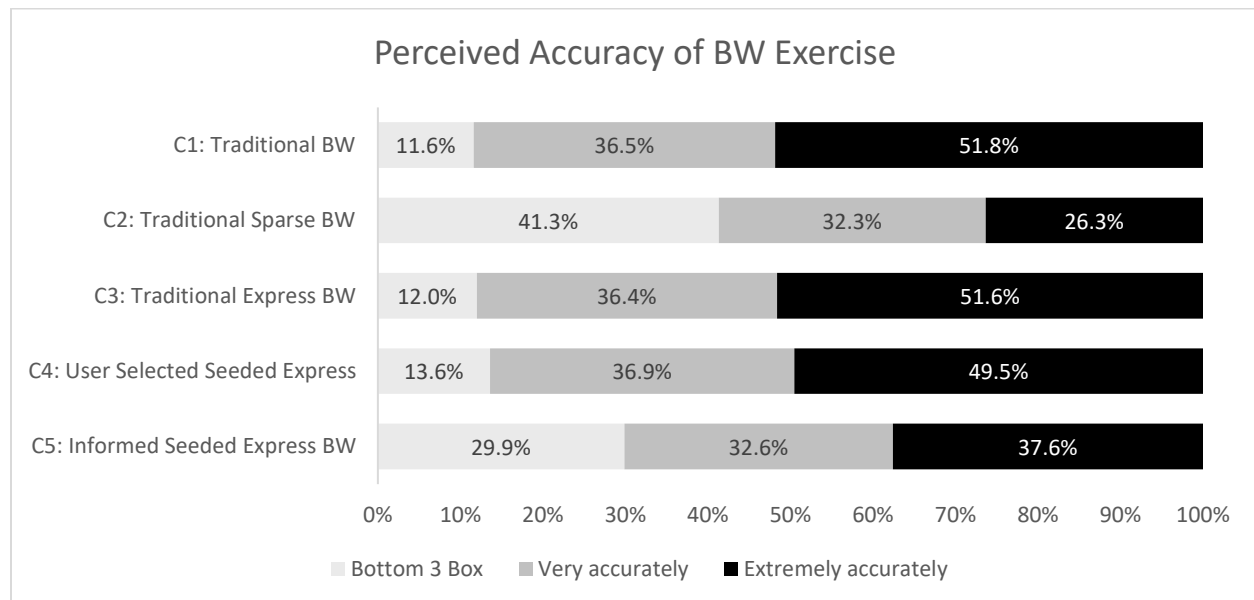
In sum, and consistent with our results with the synthetic data, we do not see much benefit for real respondents of this idea of seeding the Express BestWorst designs with fixed items seen by all respondents.

RESPONDENT BEHAVIOR AND PERCEPTIONS

While the predictive power of our models may not have worked out in our favor, do any of the approaches lead to different perceived respondent experiences when completing the different exercises each of the respondents were exposed to? We looked at several different measures to understand how the respondent experience differed across the design cells.

First, we asked respondents about their perceptions of the accuracy of the BestWorst exercise they were assigned as it pertained to their own personal preferences among the set of 40 candies. After completing the exercise, respondents were shown their personal top two and bottom two items based on on-the-fly individual-level logit estimations of their choices. We then asked them how accurately the results reflected their preferences in order to see whether any particular approach showed perceived improvement over others.

Results are shown in the bar chart below.

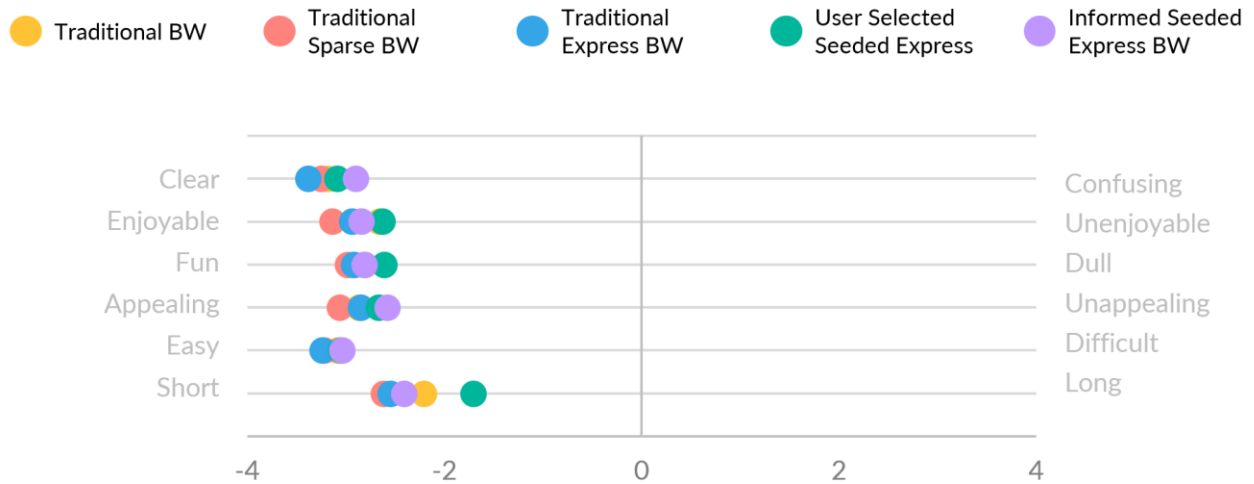


While the top two box scores were almost 90% for cells one, three, and four, they dropped to about 70% for cell five (Informed Seeded Express design) and only 60% for cell two. In other words, designs where items were shown multiple times were perceived as being more accurate than the Sparse design where each item was only shown once. This outcome is actually as expected, as it again displays a methods bias where having only one exposure to each item leads to a perceived worse fit for a particular respondent, primarily because the on-the-fly individual-level logit results from a Sparse design are less stable because of that sparseness of input information. Seeing an item only once may not accurately capture the preferences between, say, the top item from two different tasks. Once we get to actual hierarchical Bayes estimation where the individual level utilities can “borrow” from the aggregate-level preferences in the upper-level model (also known as Bayesian Shrinkage), we expect the perceptions might improve. So, we knew this pattern for the Sparse design cell would likely emerge from the data, and it does not mean that the Sparse design does not work well—we’ve shown the contrary in our Hit Rate and MAE tests above. It’s just not likely to perform well in this kind of test. The important thing to take away from these results is that the Seeded designs did not outperform the Traditional Express design in terms of perceived accuracy either.

Next we asked respondents to evaluate their own survey-taking experience in terms of how they felt about the exercise on various perception metrics. For each cell, we presented a semantic differential question with six different pairs of items covering perceptions of whether the exercise was:

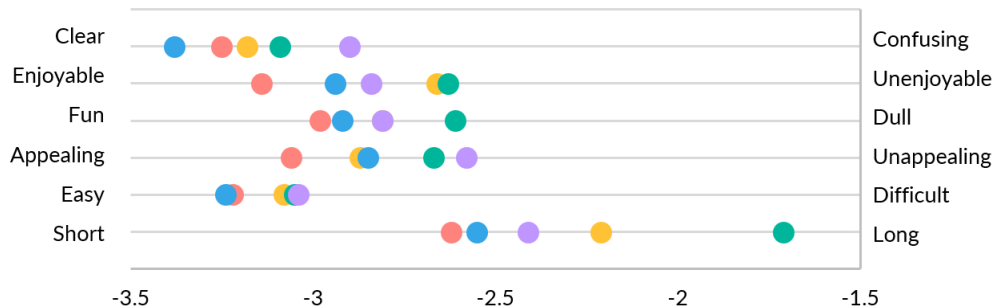
- Clear vs. Confusing
- Enjoyable vs. Unenjoyable
- Fun vs. Dull
- Appealing vs. Unappealing
- Easy vs. Difficult
- Short vs. Long

Prior to analysis, we recoded the four-point scale used to a -4, -1, 1, 4 scaling where -4=strong agreement with the statement on the left, -1= mild agreement with the statement on the left, 1 = mild agreement with the statement on the right, and 4=strong agreement with the item on the right. Just to keep you on your toes, the statements with positive connotations have negatively scaled scores and the statements with negative connotations have positively scaled scores.



Overall, you can see that regardless of design, respondents generally had strongly positive perceptions of the candy evaluation exercise. (We may have benefitted from a fun subject matter, rather than evaluating long statements about insurance benefits, no offense to any insurers who might read this paper).

If we zoom further in, we can see that the Traditional Sparse and Traditional Express designs tended to be perceived the most favorably, while the two Seeded Express designs and Traditional BestWorst design were perceived relatively worse.



ANOVA	F	p-value
Short vs. Long	10.377	<.001
Easy vs. Difficult	0.953	0.432
Appealing vs. Unappealing	3.050	0.016
Fun vs. Dull	1.756	0.135
Enjoyable vs. Unenjoyable	3.355	0.010
Clear vs. Confusing	3.165	0.013

The F test results shown above suggest that there were no strong differences in perceptions between Easy vs. Difficult nor Fun vs. Dull across the designs, but perceptions of length were highly significantly different (Traditional Sparse perceived as the shortest exercise, User Selected Seeded Express and Traditional BestWorst were perceived as the longest exercises). Given the number of screens respondents were exposed to for each of the designs, these perceptions are in line with the reality of the nature of the designs.

Once again, though, we do not see any evidence that the Seeded designs were perceived *more* positively than the traditional Express design; unfortunately, they are generally worse.

DISCUSSION

In conclusion, seeding the items in an Express BestWorst design appears to offer no real improvement over a traditional Express BestWorst design where all of the items shown are simply randomly selected from a larger pool of items. We gave the Express BestWorst designs the best chance of success by using one-half of the total items for each respondent rather than merely one-third, but having a small number of fixed items seen by all respondents did nothing to improve the in-sample or especially the out-of-sample accuracy of the predictions.

The one exception may be when there is a lot of error in the responses (i.e., high scale), as simulated data tests show higher correlation with known utilities for the seeded designs in that special case.

But it sounded like a good idea!

Sparse BestWorst designs generally outperform Express BestWorst designs, as has been shown in previous bake-off studies listed in the appendix. For large item sets with relatively simple concepts like pictures of candy, traditional BW designs do perform quite well, and the increased length is not perceived so poorly that the overall experience is viewed negatively. As the number of items in the set grows, or as the complexity of the items being studied increases, our previous work (Godin et al., 2023) suggests that Sparse designs will start to outperform traditional full designs, especially when covariates are included in the estimation of utilities.

SUMMARY

Clients sometimes balk at the idea of simply showing each item in a large design just once to each respondent, even though the Sparse design approach has over and over been shown to perform well, and relatively better than other techniques used in the presence of large design sets. While there is something perceptually comforting about having items being evaluated multiple times, when the full set of items is not exposed to each respondent, the resulting utility estimates consistently suffer especially when it comes to out-of-sample prediction. Getting some read on every item from every respondent, even if the individual signal is fairly weak, seems to better inform the model about the full gamut of preferences that may exist in the marketplace at large.

If getting highly-accurate reads of each individual's true preferences is the paramount goal of your research, provided that the item set is not too large (say, 40 items or less) or the items themselves do not require lengthy processing by respondents, a standard BestWorst design is still your best bet.

However, all of the designs we tested did produce preference data that was very highly correlated, so if you or your client have a preference for an Express BestWorst design structure despite the evidence in favor of Sparse designs, the results won't completely lead you astray either.



Thomas Eagle



Jon Godin



Megan Peitz

REFERENCES

- Chrzan, Keith and Bryan Orme (2022), “Real-Time Detection of Random Respondents in MaxDiff,” Sawtooth Software Research Paper Series (available at www.sawtoothsoftware.com/resources/technical-papers).
- Chrzan, Keith and Megan Peitz (2019), “Best-Worst Scaling with Many Items,” *Journal of Choice Modeling*, Vol. 30, March 2019, pp 61–72. (See <https://www.sciencedirect.com/science/article/pii/S1755534517301355?via%3Dihub>)
- Cohen, Steven H. (2003), “Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation.” 2003 Sawtooth Software Conference Proceedings, pp 61–74, Provo, UT.
- Godin, Jon, Abby Lerner, Megan Peitz, and Trevor Olsen (2023), “How Sparse is Too Sparse? Testing Whether Sparse MaxDiff Designs Work Under More Extreme Conditions,” 2023 Analytics & Insights Summit Proceedings, 11–29.
- Orme, Bryan (2019), “Sparse, Express, Bandit, Relevant Items, Tournament, Augmented, and Anchored MaxDiff—Making Sense of All Those MaxDiffs!,” Sawtooth Software Research Paper Series (available at www.sawtoothsoftware.com/resources/technical-papers).
- Serpetti, M., Ce. Gilbert, and M. Peitz (2016), “The Researcher’s Paradox: A Further Look at the Impact of Large-Scale Choice Exercises.” 2016 Sawtooth Software Conference Proceedings, pp 147–162, Provo, UT.
- Wirth, Ralph and Annette Wolfrath (2012), “Using MaxDiff to Evaluate Very Large Sets of Items.” 2012 Sawtooth Software Conference Proceedings, Provo, UT.