How Sparse Is Too Sparse? Testing Whether Sparse MaxDiff Designs Work Under More Extreme Conditions

Jon Godin
Abby Lerner
Megan Peitz
Trevor Olsen
Numerious Inc.

Executive Summary

Clients often want to test many items using MaxDiff. Previous research shows that Sparse MaxDiff is a valid technique for testing these conditions; however, these designs typically include many alternatives per task (5 or more). What happens when the items are extremely wordy or long? Having to read through many wordy alternatives per screen across ten or twenty screens seems to be quite burdensome. But with triplets or, even worse, paired comparisons, we get much less information from each task.

In our study, we utilized a set of 30 statements about the environment, each containing between 250–300 characters, in order to ascertain which environmental issues were more or less urgent to solve now rather than leave for future generations, testing across five different design conditions: **Traditional MaxDiff** (4 statements per task, 23 tasks), **Traditional Sparse MaxDiff** (4 statements per task, 8 tasks), **Express MaxDiff** (4 statements per task, 12 tasks, 15 of 30 statements randomly selected per respondent), **Extreme Sparse Pairs** (2 items per task, 15 tasks), and **Extreme Sparse Triplets** (3 items per task, 10 tasks). We find that Traditional MaxDiff does a better job of capturing individual preferences accurately but fares worse than other methods when making out-of-sample predictions and makes for a more painful respondent experience with higher dropout rates, higher disqualification rates, and higher inducement to cheat while answering (i.e., answer randomly to finish the task more quickly). On the other hand, Traditional Sparse MaxDiff or a best-only Paired Comparison exercise provide both a much better respondent experience and better out-of-sample rank-order predictions, especially when including covariates during HB utility estimation.

Background and Motivation

Since the time when Steve Cohen introduced Maximum Difference Scaling (MaxDiff) to the greater Sawtooth community at the 2003 Sawtooth Software Conference, MaxDiff has become a popular approach to uncovering respondent preferences among a set of items. Researchers have used MaxDiff to determine preferences for things such as advertising claims, product benefits, product messaging, images, product names, brands, features, packaging options, political voting preferences, etc.

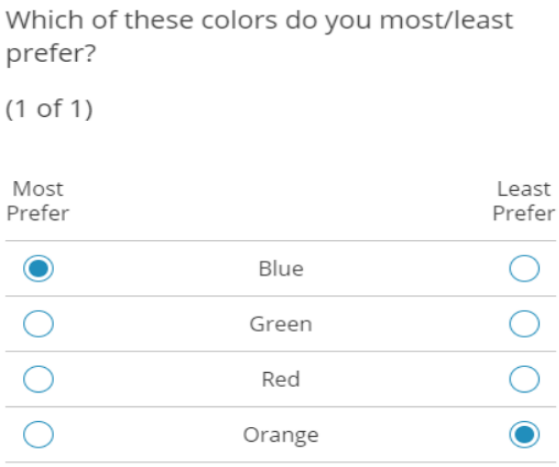
In a typical MaxDiff exercise, respondents are shown between 2–6 items at a time, and are asked to indicate which item is best and which item is worst among the set shown (different framing can be used such as most/least motivating, most/least appealing, and others). The task is repeated many times, showing a different set of items in each task, typically using enough screens/tasks so that each item is seen by each respondent at least three times. The resulting model, using Hierarchical Bayes (HB) to estimate individual-level utilities then transforming the data into ratio-scaled probability or importance scores that sum to 100 across the items, provides the ability to understand both rank order of preference among the items as well as distances between the items (i.e., an item with a score of 10 is 2x more important or more preferable than an item with a score of 5).

As the appetite for MaxDiff grew, so did client requests to include more and more items in the set to be evaluated. With more items, many more tasks would be necessary for each item to be seen three times, but that could be burdensome for respondents. This led researchers such as Wirth and Wolfrath (2012) to test more sparse data collection methods, either using only a subset of items for each respondent (called Express MaxDiff), or still using all items, but only showing each item once to each respondent (termed Sparse MaxDiff). In a Sparse MaxDiff design, then, for 60 items you might show 15 sets of 4 items, or for 120 items, you might show 24 sets of 5 items; the important part is that each item is shown about once per respondent. Despite expectations to the contrary, Wirth and Wolfrath found that Sparse MaxDiff designs outperformed Express MaxDiff designs.

Chrzan and Peitz (2019) built upon this research with a study attempting to validate Wirth and Wolfrath's findings. They found that we can run an HB multinomial logit with fairly similar estimation when each item is shown just 1x per respondent compared to 3x per respondent (albeit with less precision at the individual level and more Bayesian smoothing).

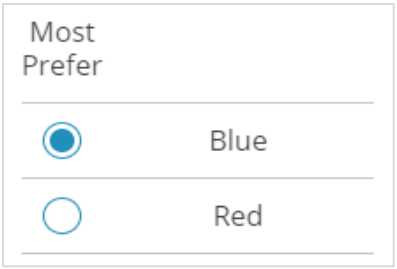
These and other examples of prior research on Sparse MaxDiff (such as Serpetti et al., 2016) all used lists of items with relatively short statements or few total characters, such as "Is made with natural ingredients" or "Has a creamy texture." However, more and more we are being asked by our clients to test very long, high-character count statements or messages. For these exercises, do we still need to display the MaxDiff tasks in quads or quints, or will triplets or even pairs be sufficient?

Theoretically, quads and quints appear to provide much more information than tasks with fewer items. For example, in the example task below where we are trying to elicit color preferences, from just two clicks we are able to ascertain five preference relationships: Blue beats Red, Green, and Orange; Red beats Orange; and Green beats Orange:



The only pair we learn nothing about from this task is whether Red beats Green or Green beats Red.

However, in the following task examples with only 3 or 2 items included, we seem to gain much less information. For triplets, we learn that Blue beats Red and Green, and Red beats Green, and for pairs we only learn that Blue beats Red, but in neither case do we learn about preferences for the other colors in the design.



Do we still have enough information in these smaller tasks to get stable preference estimates given that we're still only showing each item once to each respondent using a Sparse design approach?

Building on that issue, do these sparser approaches work better or worse when you have lengthy lists of long statements, such as this example which contains 296 characters:

> *"The average person produces 4.3 pounds of waste per day, and the U.S. accounts for 220 million tons of waste per year. This creates an environmental threat, as non-biodegradable trash gets dumped in the water, while waste from landfills generates methane, a greenhouse gas causing global warming."*

With long statements, and a lot of them (30 or more), a full MaxDiff exercise showing each item to each respondent at least three times in sets of four or five items just feels very burdensome. Do we risk burning out respondents, leading to poor data quality or inducing higher dropout rates? Alternative designs—Paired Comparisons, or MaxDiff tasks shown in triplets, or

possibly even a reduced Express MaxDiff-style design where not all items are seen by each respondent—all seem like they could make things more manageable for respondents, but would the results suffer when we get less information per task? This is what we sought to find out.

Current Research Plan

In order to study the combination of high-character-count statements in a non-traditional MaxDiff exercise, we decided to focus on trying to learn people's preferences regarding which environmental concerns they believe should be addressed now rather than pushing them off for following generations to solve.

We used a combination of old-school web searches, ChatGPT queries, and human collating and editing of these various sources into a cohesive and broad list of 30 environmental concerns, each of which ranged in length from 251 to 298 characters:

#	Statement	Num. Characters
1	Deforestation means clearing of green cover and making that land available for residential, industrial or commercial purposes. Forests cover 30% of the land, but every year tree cover is lost. Loss of forests leads to loss of biodiversity, carbon sequestration, and disruption of local communities.	298
2	Plastic pollution in oceans harms marine life and ecosystems, and can enter the human food chain. Oceans have become a giant waste dump for plastic. Unregulated disposal of waste and other materials into the ocean degrades marine and natural resources, and poses human health risks.	282
3	Water is vital for human, animal and plant survival, but water scarcity currently affects more than 40% of the world population. Growing population and industrialization are putting pressure on freshwater resources, impacting agriculture, industry, and leading to economic losses.	280
4	Air pollution in cities causes respiratory illnesses and other health problems, and contributes to climate change. Heavy metals, nitrates and plastic are among the toxins responsible for pollution, with industry and motor vehicle exhaust listed as the No. 1 pollutant.	268
5	Ocean acidity has increased in the last 250 years, but by 2100, it may shoot up by 150%. Carbon emissions are causing this impact, with 25% of total atmospheric CO2 being produced by humans. It affects ocean life and the industries that depend on it, such as fishing and tourism.	279
6	Food security around the world depends upon what condition the soil is in to produce crops. 12 million hectares of farmland is degraded each year, largely due to erosion, overgrazing, overexposure to pollutants, monoculture planting, soil compaction, and land-use conversion.	275
7	The intensive agriculture practices used to produce food have damaged the environment with the use of chemical fertilizer, pesticides and insecticides. Overuse of chemicals in agriculture harms human health too, and can lead to the development of pesticide-resistant pests.	273
8	Overfishing has a detrimental effect on natural ecosystems and leads to an imbalance of ocean life. It not only causes fishing fleets to migrate to new waters, depleting fish stocks, but also has negative effects on coastal communities that rely on fishing to support their living.	281
9	The ozone layer is an invisible layer of protection around the planet that protects life on earth from the sun's harmful UV rays. Toxic gases are creating a hole in the ozone layer, and the depletion of this layer can lead to increased skin cancer or other health problems.	273
10	There is enough evidence to show that sea levels are rising, and the melting of Arctic ice caps and glaciers worldwide, is a major contributor. Over time, the melting of polar ice caps could lead to extensive flooding, contamination of drinking water and major changes in ecosystems.	283
11	Genetic modification of food using biotechnology is called genetic engineering. It can cause environmental problems, as an engineered gene may prove toxic to wildlife. The genetic engineering of food may also cause allergic reactions and increase resistance to antibiotics for humans.	284
12	Urban sprawl refers to population migration from high-density urban areas to low-density rural areas, causing plants and animals to be displaced from their natural environment. It leads to a decline in biodiversity, and has negative effects on the social life and economy of cities.	282
13	The average person produces 4.3 pounds of waste per day, and the U.S. accounts for 220 million tons of waste per year. This creates an environmental threat, as non-biodegradable trash gets dumped in the water, while waste from landfills generates methane, a greenhouse gas causing global warming.	296

14	People around the world use so many natural resources that we would need almost 1.5 Earths to cover all our needs. The increased use of these natural resources has led to industrialization and air pollution. Over time, natural resource depletion will lead to an energy crisis.	276
15	Human activity is leading to the extinction of species and habitats and loss of biodiversity. Ecosystems are in danger when any species' population is decimating. More than 500 species of land animals are on the brink of extinction and are likely to be lost within 20 years.	274
16	The increase of global warming from CO2 emissions is accelerating climate change. Climate change threatens the survival of millions of people, plants and animals by causing more extreme, frequent meteorological events, like droughts, fires and floods.	251
17	Illegal fishing is threatening wildlife. A shocking 640,000 tons of abandoned, lost, or otherwise discarded fishing gear is left in the world's oceans each year, which entangles and kill around 136,000 turtles, whales, seals, birds, and other sea animals.	255
18	Noise pollution is regular exposure to elevated sound levels that leads to adverse effects in humans or other living organisms. Exposure to loud noise can cause hearing loss, high blood pressure and heart disease in humans, while also negatively impacting the health and well-being of wildlife.	294
19	The world's food system is responsible for up to one third of all human-caused greenhouse gas emissions. The conventional agriculture industry has an enormous carbon footprint; it not only covers a vast amount of land, but also consumes a vast amount of freshwater.	265
20	Nuclear reactions can result in widespread contamination in air and water, aside from the loss of human life. Though nuclear reactors do not generate air pollution or carbon dioxide while operating, radioactive waste is toxic. It can cause cancer and damage to the immune system.	279
21	Cobalt is a key component of battery materials that power electric vehicles. Cobalt mining, however, is associated with many environmental and social issues. Mining regions have high radioactivity levels, and dust from pulverized rock is causing breathing problems for local communities.	287
22	Wetlands provide vital ecosystem services, such as water purification, flood control, and wildlife habitat. Wetland loss can add stress to remaining wetlands, and can also decrease habitat, landscape diversity, and connectivity among aquatic resources.	252
23	Non-native species are organisms not found naturally in an area, but are introduced as the result of human activities. Invasive non-native species are capable of causing extinctions of native plants and animals, competing with these organisms for limited resources and altering their habitats.	293
24	The world's population is more than three times larger than it was in the mid-twentieth century. Population growth not only affects food security, but also the livelihoods of farmers. As population grows, so does the demand for food, putting strain on agriculture and natural resources.	286
25	Fracking or extractive industry consists of the people, companies, and activities involved in removing oil, metals, coal, stone, and other materials from the ground. Such industry practices can cause habitat destruction, pollution, and disruption of local communities and their livelihoods.	290
26	Acid rain can be caused due to the combustion of fossil fuels or erupting volcanoes or rotting vegetation, which releases sulfur dioxide and nitrogen oxide into the atmosphere. It can also be caused by human activities. It can impact human health, wildlife, and aquatic species.	278
27	Desertification is the process by which vegetation in drylands, such as grasslands or shrublands, decreases and eventually disappears. Desertification can lead to loss of biodiversity and displacement of local communities, as well as increase the risk of zoonotic diseases.	273
28	Through the emissions from combustion of fossil-derived fuels, transportation systems contribute to degraded air quality, as well as a changing climate. Transportation also leads to noise pollution, water pollution, and affects ecosystems through multiple direct and indirect interactions.	289
29	The global demand for fashion and clothing now accounts for 10% of global carbon emissions, becoming an increasing problem. In addition to greenhouse gas emissions, textile dyeing and microplastics from various materials pollute wastewater and discarded clothing ends up in landfills.	284
30	The number of natural disasters that cost over a billion dollars has increased over the last forty years, rising from an average of 3 per year in the 1980s to 13 per year during the 2010s. Not only are natural disasters occurring more frequently, their average cost and death toll is up as well.	295

For this research, respondents would be shown only the full statements as listed above, with no use of any simplifying techniques commonly used in practice such as bolding, highlighting, or italicizing key words, providing shorter definitions on-screen with hover-overs of the full definitions, or the like. We purposefully did not want to make this easy, and perhaps sought to make it a bit painful for respondents. We think you'll agree that even reading through the statements above once is a lot to take in.

Our research design utilized five design cells, each varying either the frequency that each item would be shown to a given respondent (1x to 3x), the number of statements included per task (2, 3, or 4), and/or the number of statements included per respondent (either a random selection of 15, or the full set of 30). The specific cells we tested were:

Cell #	MaxDiff Design Description	N Size	# Tasks
1	Traditional MaxDiff, 4 items/task, each shown 3x	302	23
2	Traditional Sparse MaxDiff, 4 items/task, each shown 1x	303	8
3	Express MaxDiff, 4 items/task, 15 items per respondent, each shown 3x	306	12
4	Extreme Sparse Pairs, 2 items/task, each shown 1x	303	15
5	Extreme Sparse Triplets, 3 items/task, each shown 1x	301	10

All designs used 200 versions. It's worth nothing that for the Express MaxDiff cell, we used a design that included half of the items being shown to each respondent, which has not always been the case in earlier research, in order to give the Express approach a better chance of succeeding. Respondents in that cell would receive a randomized sampling of 15 of the 30 items, with a different randomization used for each of the 200 versions.

The study was programmed and hosted using Sawtooth Software's Lighthouse Studio. The designs for Cells 1, 2 and 3 were created within Lighthouse Studio, while the designs for Cells 4 and 5 were created using Numerious's Julia-based designer in order to ensure perfect level balance (1x) in each version of the design.

Screens for Cells 1–3 would look similar, displaying four items per task and asking respondents to indicate the most and least problematic environmental concern among the set shown. Cell 4 would only display two statements per screen, asking respondents to only indicate which statement was the most problematic, while Cell 5 would show three per screen and again ask both most and least problematic statements be identified.

In addition to the main design for each cell, we included two fixed holdouts for in-sample testing using the same structure as the main design of the cell. These holdouts were created via two-task, 1 version supplemental designs using Lighthouse Studio's MaxDiff designer. For Cell 3 holdouts (Express MaxDiff), we did not use the Serpetti et al. approach of creating an "Express Unique Anchor" holdout only showing items that a given respondent would have personally evaluated in the exercise. Therefore, the fixed holdouts for this cell are knowingly somewhat problematic, since there is no guarantee that a given respondent saw any of the four statements in each holdout during their MaxDiff exercise due to the randomization of the items entering each respondent's design.

Screenshots of each of the respondent tasks for the five cells are shown below:

Cell 1: Traditional MaxDiff

Considering these four statements, which is the **most problematic** to leave for future generations to address instead of trying to solve now? Which is **least problematic**?

(1 of 25)

Most Problematic		Least Problematic
○	Air pollution in cities causes respiratory illnesses and other health problems, and contributes to climate change. Heavy metals, nitrates and plastic are among the toxins responsible for pollution, with industry and motor vehicle exhaust listed as the No. 1 pollutant.	○
○	Fracking or extractive industry consists of the people, companies, and activities involved in removing oil, metals, coal, stone, and other materials from the ground. Such industry practices can cause habitat destruction, pollution, and disruption of local communities and their livelihoods.	○
○	Nuclear reactions can result in widespread contamination in air and water, aside from the loss of human life. Though nuclear reactors do not generate air pollution or carbon dioxide while operating, radioactive waste is toxic. It can cause cancer and damage to the immune system.	○
○	Deforestation means clearing of green cover and making that land available for residential, industrial or commercial purposes. Forests cover 30% of the land, but every year tree cover is lost. Loss of forests leads to loss of biodiversity, carbon sequestration, and disruption of local communities.	○

Cell 2: Traditional Sparse MaxDiff

Considering these four statements, which is the **most problematic** to leave for future generations to address instead of trying to solve now? Which is **least problematic**?

(1 of 10)

Most Problematic		Least Problematic
○	There is enough evidence to show that sea levels are rising, and the melting of Arctic ice caps and glaciers worldwide, is a major contributor. Over time, the melting of polar ice caps could lead to extensive flooding, contamination of drinking water and major changes in ecosystems.	○
○	Fracking or extractive industry consists of the people, companies, and activities involved in removing oil, metals, coal, stone, and other materials from the ground. Such industry practices can cause habitat destruction, pollution, and disruption of local communities and their livelihoods.	○
○	Wetlands provide vital ecosystem services, such as water purification, flood control, and wildlife habitat. Wetland loss can add stress to remaining wetlands, and can also decrease habitat, landscape diversity, and connectivity among aquatic resources.	○
○	Overfishing has a detrimental effect on natural ecosystems and leads to an imbalance of ocean life. It not only causes fishing fleets to migrate to new waters, depleting fish stocks, but also has negative effects on coastal communities that rely on fishing to support their living.	○

Cell 3: Express MaxDiff

Considering these four statements, which is the **most problematic** to leave for future generations to address instead of trying to solve now? Which is **least problematic**?

(1 of 14)

Most Problematic		Least Problematic
○	Ocean acidity has increased in the last 250 years, but by 2100, it may shoot up by 150%. Carbon emissions are causing this impact, with 25% of total atmospheric CO2 being produced by humans. It affects ocean life and the industries that depend on it, such as fishing and tourism.	○
○	Overfishing has a detrimental effect on natural ecosystems and leads to an imbalance of ocean life. It not only causes fishing fleets to migrate to new waters, depleting fish stocks, but also has negative effects on coastal communities that rely on fishing to support their living.	○
○	Genetic modification of food using biotechnology is called genetic engineering. It can cause environmental problems, as an engineered gene may prove toxic to wildlife. The genetic engineering of food may also cause allergic reactions and increase resistance to antibiotics for humans.	○
○	Food security around the world depends upon what condition the soil is in to produce crops. 12 million hectares of farmland is degraded each year, largely due to erosion, overgrazing, overexposure to pollutants, monoculture planting, soil compaction, and land-use conversion.	○

Cell 4: Sparse Paired Comparisons

Considering these two statements, which is the **most problematic** to leave for future generations to address instead of trying to solve now?

(1 of 17)

Most Problematic	
○	Acid rain can be caused due to the combustion of fossil fuels or erupting volcanoes or rotting vegetation, which releases sulfur dioxide and nitrogen oxide into the atmosphere. It can also be caused by human activities. It can impact human health, wildlife, and aquatic species.
○	Overfishing has a detrimental effect on natural ecosystems and leads to an imbalance of ocean life. It not only causes fishing fleets to migrate to new waters, depleting fish stocks, but also has negative effects on coastal communities that rely on fishing to support their living.

Click the 'Next' button to continue...

Cell 5: Sparse Triplet MaxDiff

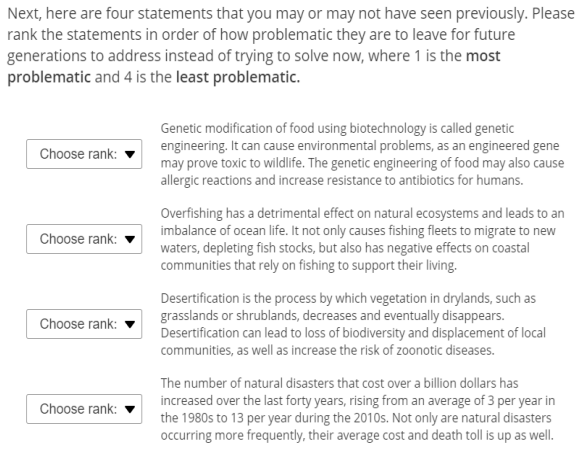
Considering these three statements, which is the **most problematic** to leave for future generations to address instead of trying to solve now? Which is **least problematic**?

(1 of 12)

Most Problematic		Least Problematic
○	Plastic pollution in oceans harms marine life and ecosystems, and can enter the human food chain. Oceans have become a giant waste dump for plastic. Unregulated disposal of waste and other materials into the ocean degrades marine and natural resources, and poses human health risks.	○
○	Acid rain can be caused due to the combustion of fossil fuels or erupting volcanoes or rotting vegetation, which releases sulfur dioxide and nitrogen oxide into the atmosphere. It can also be caused by human activities. It can impact human health, wildlife, and aquatic species.	○
○	The average person produces 4.3 pounds of waste per day, and the U.S. accounts for 220 million tons of waste per year. This creates an environmental threat, as non-biodegradable trash gets dumped in the water, while waste from landfills generates methane, a greenhouse gas causing global warming.	○

Click the 'Next' button to continue...

In addition to the cell-specific in-sample holdouts, we also created two universal fixed holdout questions using a ranking task. The same two ranking tasks were shown to all respondents, regardless of design cell. Within each ranking task, we asked respondents to select a ranking for each of the four statements shown, where 1 = the most problematic issue, and 4 = the least problematic issue. An example task is shown below:



In all cases, the holdouts were not used in model estimation. Instead, estimated utilities from each of the design cells would be used to predict holdout choices for in-sample tasks, and item rank orders for the out-of-sample ranking tasks. The ranking questions always included four items per screen, which potentially could bias results towards those cells also showing four items per task (i.e., Cells 1–3).

Fieldwork was conducted between February 10–17, 2023, using Prodege's peeq marketplace sample among respondents age 18+, with no other screening being used. Collected data was cleaned for speeding (< 1/3 median completion time) as well as an age mismatch (stated age asked early in survey vs. year of birth asked at the end of the survey, screening out those with a mismatch of 2 years or more). Additional data collected included gender, income, SASSY segment[1], home ownership, home area and state of residence, clean energy usage, attitudes towards climate and the environment, and political affiliation.

As an additional note on data cleaning, we did not use any on-the-fly Root Likelihood (RLH) comparisons vs. dummy respondents to clean bad cases while in field. While we like to use this approach in general practice, here we wanted to test whether any of the approaches naturally caused bad respondent behavior, so we didn't want to screen people out prematurely. In addition, for the sparse approaches we tested, the RLH test isn't really reliable with items being seen less than 3x per respondent, so we couldn't apply it consistently here even if we wanted to include on-the-fly quality testing.

[1] SASSY segments were derived from the Yale Program on Climate Change Communications Six Americas Super Short Survey (SASSY), found here: https://climatecommunication.yale.edu/visualizations-data/sassy/

For each cell, we estimated two Hierarchical Bayes models: one with no covariates, and one including gender, income, age generation, SASSY segment, and political party affiliation as covariates. Each model utilized 20,000 burn-in and 20,000 saved iterations, otherwise using standard Lighthouse Studio estimation defaults.

Finally, we also estimated an overall model using Sawtooth Software's stand-alone CBC/HB module by collapsing the .cho (choice) files from each of the five cells into one single file, also estimating the model twice, once without and once with the covariates listed above.

Analysis of Results

In-Sample Holdouts

First, we assess in-sample validity by computing individual-level hit rates and aggregate-level Mean Absolute Errors (MAEs) when comparing actual holdout choices to those predicted from the estimated utilities for each cell. These were computed for both Best and Worst choices, but for space-saving reasons we only show the overall averages across these in the table below.

In-Sample Hit Rates (Higher is Better)

Overall Hit Rates	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse MaxDiff Triplets
No covariates	45.9%	49.8%	46.9%	76.9%	54.2%
With covariates	46.2%	48.4%	45.7%	74.6%	53.7%
Difference	+0.3%	-1.4%	-1.2%	-2.3%	-0.5%

Both without and with covariates, Sparse Quads (Cell 2) achieve the highest hit rates among the 4-item holdout methods (Cells 1–3), and results otherwise seem reasonable. Obviously, with either pairs or triplets it's easier to get a hit than it is with quads, as the results reflect.

For predicting individual-level choices, adding covariates to the model doesn't seem to help and in fact for most cells slightly hurts the predictions, though the differences aren't operationally meaningful.

In-Sample Mean Absolute Errors (Lower is Better)

Average MAEs	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse MaxDiff Triplets
No covariates	1.7%	5.3%	3.8%	7.2%	3.5%
With covariates	1.8%	5.2%	3.5%	1.3%	4.0%
Difference	+0.1%	-0.1%	-0.3%	-5.9%	+0.5%

Moving on to MAEs, though it is sometimes the practice of academics and practitioners to tune the model exponent for each cell to minimize the within-cell MAEs, we did not take that step so the MAEs shown above are "natural." Results-wise, we see that the full traditional MaxDiff design (Cell 1) achieves the lowest MAEs in-sample when no covariates are included in the model; Express MaxDiff also performs relatively well here given its methods bias disadvantage.

However, unlike with Hit Rates, the inclusion of covariates generally helped lower the MAEs, but only slightly in most cases, except for the Sparse Pairs cell which saw dramatic improvement. We surmise that for the Sparse Pairs the covariates are helping to reel in more extreme preferences at the individual level, leading to improved predictions of whether a given item is better than another item without overstatement.

Out-of-Sample Holdouts

Although ensuring in-sample validity is important, we feel that achieving a better ability to predict out-of-sample choices or preferences is really the gold standard for model comparisons. Here, rather than trying only to predict the overall out-of-sample rankings (considering the rankings of all 4 items at once) for each holdout ranking task, which is a very high hurdle to clear accuracy-wise, we cycled through all of the different iterations of rankings that could be derived from the data:

- **Pairs**—for any given pair in the rankings holdout, can we predict the relative ranking correctly? (18 pairs evaluated)
- **Triples**—for any given set of 3 items in the rankings holdout, can we predict the relative ranking correctly? (8 triples evaluated)
- **Quads**—for the whole set of 4 items in each ranking holdout, can we predict the relative ranking correctly? (2 quads evaluated)

In the tables that follow, we computed a weighted average across all of these splits for each cell for easier comparisons. For the Combined Model, the results represent the average across all cells. Once again, we look at both Hit Rates and MAEs for each of the methods.

Ranking Hit Rates (Higher is Better)

Ranking Hit Rates	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse MaxDiff Triplets	Combined Model
No covariates	68.0%	63.8%	65.3%	61.0%	60.4%	64.7%
With covariates	68.2%	62.7%	61.4%	57.9%	60.2%	64.5%
Difference	+0.2%	-1.1%	-3.9%	-3.1%	-0.2%	-0.2%

As we might expect, the Sparse Cells (2, 3, and 4) perform slightly worse on hit rates than the methods where each item is shown at least 3 times to each respondent, though all methods are roughly comparable to the combined model benchmark. In this case, all cells saw ranking holdout tasks with four items each, so we wouldn't expect the Pairs or Triples to outperform the quads as we saw with the cell-specific holdouts shown earlier.

While Express MaxDiff has performed poorly in other bakeoff tests, it does surprisingly well here where we included a larger (50%) sampling of the full set of items.

As we saw for in-sample holdouts, hit rates for the rankings holdouts are generally slightly worse when covariates are included in the HB estimation.

Out-of-Sample MAEs (Lower is Better)

Ranking Aggregate MAEs	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse MaxDiff Triplets	Combined Model
No covariates	7.2%	4.5%	11.7%	7.9%	7.9%	7.3%
With covariates	6.3%	3.1%	9.1%	4.2%	5.5%	6.5%
Pct.-point improvement	-0.9	-1.4	-2.6	-3.7	-2.4	-0.8
% Reduction in Error	-12.5%	-31.1%	-22.2%	-46.8%	-30.4%	-11.0%

For out-of-sample MAEs, Sparse Pairs and Triplets perform almost at par with Traditional MaxDiff, but Sparse Quads achieved the lowest error rate without the presence of covariates.

For the ranking holdouts, we observe marked improvement in out-of-sample predictions when using covariates across all cells; Sparse Quads (Cell 2) still perform best, but the Sparse Pairs (Cell 4) improved the most when covariates are added to the model, nearly halving the error rate achieved without covariates, and reducing the average error rate to be much closer to the overall-leading Cell 2.

Yet again, it is in the out-of-sample predictions where we continue to see Express MaxDiff suffer relative to the other methods tested.

Importance Score Comparisons

To assess the consistency of the estimated importance (probability) scores across the cells, we ran correlations of the results for each pair of test cells as well as against the overall combined model. In the table below, which shows results for the models estimated without covariates, we see that the correlations across methods are strong, with all correlations > 0.9. Cells 1 and 3 have the highest correlation with the overall model. Though Sparse Pairs (Cell 4) have the lowest correlations with other cells, they remain relatively high.

Correlations of Importance Scores by Cell

	Cell 1: Traditional MaxDiff	Cell 2: Traditional Sparse MaxDiff	Cell 3: Express MaxDiff	Cell 4: Sparse Pairs	Cell 5: Sparse MaxDiff Triplets	Combined Model
C1: Traditional MaxDiff	1.000	0.949	0.948	0.917	0.925	0.983
C2: Sparse Quads		1.000	0.938	0.915	0.911	0.969
C3: Express MaxDiff			1.000	0.903	0.961	0.978
C4: Sparse Pairs				1.000	0.902	0.949
C5: Sparse Triplets					1.000	0.966
Combined Model						1.000

Comparing the importance scores themselves across cells (again using the data from the models without covariates), we are comforted to see that the top 2 items are the same across all cells (though the order of preference is flipped for the Sparse Pairs Cell 4), and the bottom item is the same across all cells. The relative story about the importance of the various environmental concerns is otherwise very similar across cells, with no indications of items jumping up or falling down dramatically for any cell vs. the others.

Mean Item Importance Scores

Item	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Combined Model
03 Water scarcity	6.44	6.73	6.65	5.18	5.54	6.02
02 Plastic pollution	5.36	5.58	5.76	5.23	5.12	5.36
16 Global warming	4.83	4.45	5.55	4.75	4.87	4.77
04 Air pollution in cities	4.30	5.06	4.92	4.74	4.83	4.69
01 Deforestation	4.57	5.20	4.83	4.74	4.18	4.56
06 Soil condition	4.03	4.72	4.92	3.95	4.73	4.35
10 Melting of Arctic ice caps and glaciers	4.36	3.54	3.71	4.09	4.39	4.05
09 Hole in the ozone layer	4.15	3.75	4.57	3.56	4.19	4.04
13 Waste	3.98	4.17	3.76	4.47	3.80	4.00
15 Extinction of species and habitats	3.91	3.69	3.26	4.40	3.67	3.76
14 Increased use of natural resources	3.80	3.71	3.42	3.78	3.62	3.76
30 Natural disasters	3.46	3.44	4.03	3.33	4.14	3.63
24 Population growth	3.68	4.36	3.72	3.00	3.59	3.62
28 Transportation system emissions	3.30	3.07	3.92	3.50	3.72	3.50
05 Ocean acidity	3.47	3.27	3.11	3.38	3.71	3.44
20 Nuclear reactions	3.08	2.95	3.70	3.33	3.48	3.26
19 Food system/agribusiness carbon footprint	3.49	3.15	3.18	2.78	2.89	3.25
07 Chemical fertilizer, pesticides, & insecticides	2.69	3.31	3.20	3.55	3.52	3.23
22 Wetland loss	3.21	2.68	2.87	3.18	2.31	2.94
27 Desertification	2.53	2.59	2.34	2.85	2.42	2.57
08 Overfishing	2.89	2.33	2.05	2.51	2.23	2.55
11 Genetic modification of food	2.42	2.59	2.48	2.76	2.08	2.47
26 Acid rain	2.34	1.94	2.54	2.32	2.94	2.46
25 Fracking or extractive industry	2.17	2.16	2.05	2.60	2.56	2.36
21 Cobalt mining	2.45	2.04	2.45	2.09	2.42	2.35
17 Abandoned fishing gear	2.33	2.55	1.74	2.19	2.36	2.25
12 Urban sprawl	2.13	2.43	1.59	2.71	1.90	2.15
23 Invasive non-native species	1.96	1.94	1.77	2.12	2.09	1.97
29 Fashion/clothing demand	1.45	1.53	1.32	2.10	1.47	1.56
18 Noise pollution	1.22	1.06	0.57	0.81	1.24	1.07

To sum up, when faced with designs that include a large number of high-character count items, sparse methods whether based on quads, pairs, or triples produce similar importance scores to traditional MaxDiff designs, but more importantly, seem to better predict out-of-sample preferences, especially when covariates are used in estimation.

RESPONDENT BEHAVIOR AND PERCEPTIONS

Beyond predictive accuracy and item preference consistency, we wanted to gauge respondent reactions to each of the designs, both behaviorally and attitudinally. First, we attempt to ascertain how burdensome each of the designs was for respondents by looking at respondent disqualification rates and perceptions of inducement to cheat during the MaxDiff exercise.

Based on the standard research DQ checks we used (less than one-third median time to complete and age mismatch), we removed significantly more respondents from Cell 1 (Traditional MaxDiff) than any of the other cells, but most particularly Cells 4 (Sparse Pairs) and 5 (Sparse Triplets), as shown in the table below:

	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse Triplets
Removed for DQ	**5.3%**	3.2%	2.2%	1.6%	1.6%

The p-value of the Chi-Square statistic on disqualification rate differences across cells was 0.026, so we are confident that the DQ rate differs across the cell treatments (this exceeds the 95% threshold for the statistic to be considered statistically significant). Respondents failing DQ checks were removed prior to any subsequent analysis.

We also asked two questions *after* the MaxDiff exercise to explore whether any of the designs induced bad respondent behavior, at least from a self-reported perspective. These questions were:

1. In the hope of designing better surveys for people like you, would you please tell us . . . At any point during this exercise did you feel like selecting a random answer to get through the survey faster? Now that you are done with the exercise, it's OK to be honest and you will not be penalized for answering this honestly.
2. [If yes] You mentioned that you felt like selecting random answers in order to get through this survey faster. Did you *actually* select random answers in order to finish this survey faster? Once again, you will not be penalized for your honest answer.

Results are shown in the table below. Here again we see that, at least directionally, more respondents admitted to feeling like cheating during the exercise from Cell 1 (Traditional MaxDiff) than any other cell, and all of the four-items-per-task designs (Cells 1–3) had directionally higher rates than the tasks with only pairs or triplets (Cells 4–5). Actual admitted cheating rates are relatively comparable across tasks, ranging from ~6%–8%.

	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse Triplets
Felt like cheating	22.2%	18.8%	16.3%	15.2%	15.6%
Admitted to cheating	6.3%	8.3%	5.9%	7.3%	7.6%

From a broader survey completion perspective, we looked at median total survey completion times as well as dropout rates. Timewise, the Sparse Triplets exercise required only 60% of the time needed for the full Traditional MaxDiff exercise. For the dropouts, we looked across cells

and flagged any respondents who dropped out during the respective MaxDiff exercise they were exposed to. Here we can see that 3.2x as many dropouts occurred in Cell 1 compared to Cell 4!

	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse Triplets
Median Survey Time	14.2	9.2	10.5	9.3	8.6
% of All Dropouts During Experiment	12.99%	9.60%	9.60%	3.95%	9.60%

So thus far we have at least directional evidence that respondents in the Traditional MaxDiff cell displayed more problematic survey behavior (higher DQ rates, higher dropout rates, and greater likelihood of feeling like cheating), and can confirm the survey length is much longer, which might induce these behaviors. How then did the respondents who completed the exercise feel about the experience?

To uncover these attitudes, we asked respondents a set of six semantic differential questions on a four-point scale, regarding their perceptions of whether the survey was:

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

We randomized which item was shown on the left or right as well as the order of each of the pairs during data collection. The data collected from these questions was rescaled to -4, -1, 1, 4 scaling, and items were flipped post-data collection so any "bad" items would be associated with negative scores and "good" items would be associated with positive scores. Results in the table below show that on the whole the Sparse Pairs cell (Cell 4) outperforms all of the other cells, and the Traditional MaxDiff cell (Cell 1) fares the worst by far (all results have at least directionally significant p-values from ANOVA F-tests) [italics indicate lowest score, bold text indicates highest]:

	Mean Score					ANOVA Results	
Semantic Differential Pair	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	F	p-value
Long (-) vs. Short (+)	*-0.11*	**0.87**	0.75	0.83	0.64	10.411	<.001
Difficult (-) vs. Easy (+)	*1.70*	1.92	2.00	**2.20**	1.97	1.896	0.109
Unappealing (-) vs. Appealing (+)	*1.35*	1.77	1.59	**1.85**	1.65	2.310	0.056
Dull (-) vs. Fun (+)	*0.87*	1.32	1.16	**1.63**	1.28	4.800	<.001
Unenjoyable (-) vs. Enjoyable (+)	*1.38*	1.57	1.68	**1.93**	1.59	2.422	0.047
Confusing (-) vs. Clear (+)	2.29	2.32	2.53	**2.78**	*2.22*	3.391	0.009

Lastly, we asked respondents two open-ended questions regarding what they liked and disliked about their survey experience. NLP count vectorization of the resulting comments was conducted using Python. Several patterns emerged from the data.

In terms of likes, “easy” was mentioned ~2x more frequently by Cell 4 respondents, “nothing” was mentioned 1.5x more frequently for Cell 1 than for Cells 4 or 5, and “think” (as in “made me think”) was mentioned 1.65x more frequently for Cells 2–5 than for Cell 1.

Cell 1 Traditional MaxDiff - Likes

Cell 2 Sparse Quads - Likes

Cell 3 Express MaxDiff - Likes

Cell 4 Sparse Pairs - Likes

Cell 5 Sparse Triplets - Likes

In terms of dislikes, Cell 1 respondents used “long” 2x–4x more than other cells, “hard” was mentioned 2x less in Cell 4 than any other cell, and “repetitive” was not in the top 25 most mentioned words for Cell 4, whereas it was mentioned 2x more in Cell 1 than in Cells 2, 3, or 5.

Cell 1 Traditional MaxDiff - Dislikes

Cell 2 Sparse Quads - Dislikes

Cell 3 Express MaxDiff - Dislikes

Cell 4 Sparse Pairs - Dislikes

Cell 5 Sparse Triplets - Dislikes

So, in addition to somewhat worse respondent behavior for a Traditional MaxDiff exercise composed of many very long statements, the respondents in that cell expressed fewer positive perceptions of and stronger negative feelings toward the exercise than respondents experiencing the Sparse Pairs design, and to a somewhat lesser degree, the Traditional Sparse Quads MaxDiff design.

Discussion

If we can agree that out-of-sample MAE is probably the best measuring stick, with the marked improvement when covariates are added, and factoring in respondent preference, it appears that Sparse Quads or Sparse Pairs are the best approaches for lengthy high character count MaxDiff exercises.

If you were to really need to nail individual preferences at the expense of overall accuracy, perhaps you still might consider a Traditional MaxDiff for high character count lists, but if you're willing to sacrifice a little individual-level precision for overall market accuracy, then Sparse Quads still seems to be the gold standard of Sparse methods, though Sparse Pairs is clearly preferred by respondents and fares very well as long as relevant covariates are included in the estimation. We suspect the 3x-shown methods underperform in the aggregate due to fatigue-related issues leading to response errors, which the Sparse methods don't appear to suffer from as much.

Overall Performance Ranking Scorecard

	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse Triplets
Hit Rates without covariates	1	3	2	4	5
Hit Rates with covariates	1	2	3	5	4
OOS MAE without covariates	2	1	5	3.5	3.5
OOS MAE with covariates	**4**	**1**	**5**	**2**	**3**
Respondent Preference	**5**	**2**	**3**	**1**	**4**
Overall	**5**	**1**	**4**	**2**	**3**

SUMMARY

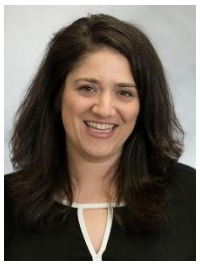
In sum, when you need to test long lists of very wordy statements, Sparse Quads or Pairs seem best. However, all approaches we tested produced importance scores that were highly correlated across cells, which is comforting.

For designs with high character count statements, Traditional MaxDiff does a better job of capturing individual preferences accurately but fares worse than other methods OOS and offers a fairly painful respondent experience with higher dropout rates, higher disqualification rates, and higher inducement to cheat being felt by respondents. Traditional Sparse MaxDiff (showing quads) or a best-only Paired Comparison exercise (with covariates included during estimation) provide both a better respondent experience and better out-of-sample rank-order predictions.

As the list size itself increases, the Pairs method may become less viable as the number of pairs required to cover all items at least once could get quite large; traditional Sparse MaxDiff should be the go-to in that case.

Express MaxDiff fared well at capturing individual-level hit rates, so if individual-level rather than market-level inferences are your goal, it might be an option, though we suggest showing at least 50% of the items to each respondent in that case.

Jon Godin



Abby Lerner

Megan Peitz

Trevor Olsen

REFERENCES

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.sciencdirect.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.

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.