Design of Experiments (DOE) Choice Design vs. Forced Ranking

Don Lifke, Research and Development Engineer, Sandia National Laboratories

Claire Syroid, Pharmacist Clinician, Walgreens Specialty Pharmacy

Background

We compare two different methods, a) the DOE Choice Design feature in JMP and b) a forced ranking methodology, for determining order of preference of items that typically only have nominal characteristics, such as taste. Force-ranking involves arranging *n* items in order of preference from 1 to *n* without any ties. A Choice Design presents items in varying pairs, requiring only that the best choice of the two be identified. (Choice designs can also be designed to present more than two profiles per choice set.)

Forced ranking is often used, but can be tedious for the judges, especially when the number of choices exceeds about five items. A Choice Design is typically easier for judges to complete. (Meyer, 2012)

A typical completed Forced Ranking judging sheet might look like this:

Rank the following items in order of preference, using ranks of 1-12 without duplication (no ties):	
Item:	**Rank (1-12):**
1	4
2	11
3	6
4	2
5	5
6	9
7	10
8	1
9	7
10	3
11	12
12	8

Figure 1: Forced Ranking Judging Sheet

A typical completed DOE Choice Design judging sheet might look like this:

Circle the preferred item (Item 1 or Item 2) in each pair:	
Item 1	**Item 2**
1	7
6	3
3	8
9	1
11	10
7	6
8	9
12	11

Figure 2: DOE Choice Design Judging Sheet

A very simple example of choice judging is "The Pepsi Challenge," in which the judges were provided an unidentified cup of Pepsi and an unidentified cup of Coke and asked to select the

one they preferred. (The cups were actually marked with Q and M.)

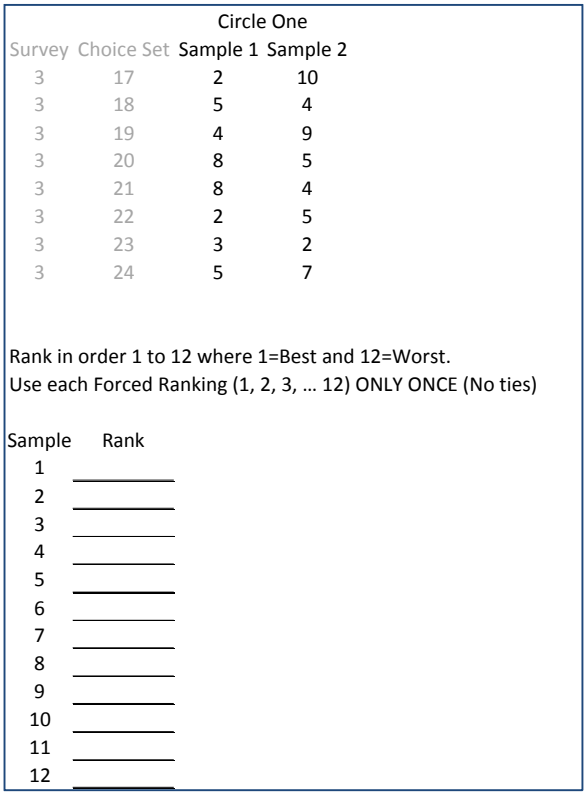
Methodology

Discrete Choice

In order to compare the two methodologies, a sufficient data set was necessary. Time and resources could be consumed creating experiments, such as employee ranking for Performance Management Forms (PMF), or vendor selection for weapons hardware, but these would provide other difficulties such as personally identifiable information and security issues. Since the intent is to compare two judging methods and not to optimize employee PMF rankings or vendor selection, it was decided to take full advantage of an upcoming event – an evening wine tasting gathering. This would provide a very nice free data set that could be understood by most people, while serving the purpose of comparing two judging methodologies. The judges were not paid, and they supplied the samples! It doesn't get much better than that for collecting experimental data. All of the judges were seasoned wine enthusiasts.

The event involved twelve different Oregon Pinot Noir wines. There were no other restrictions placed on the samples provided by the attendees, other than they needed to be Pinot Noir wines from Oregon. All twelve entries were different; there were no duplicates in the experiment. Attendees were asked to bring two bottles of the same wine they chose, but the tasting did not require more than the initial bottle for each sample. It is well-understood by wine enthusiasts that there can be a significant difference even from bottle to bottle of the same wine.

Samples were wrapped in aluminum foil in an identical manner. Labels were not visible to the judges. The wrapping was not performed by the same person as the sample numbering. This maximized anonymity of the samples. Attendees were handed judging sheets that required them to force-rank the twelve samples, and to also choose between eight pairs presented to them. Each survey was unique. A sample judging sheet follows.



		Circle One	
Survey	Choice Set	Sample 1	Sample 2
3	17	2	10
3	18	5	4
3	19	4	9
3	20	8	5
3	21	8	4
3	22	2	5
3	23	3	2
3	24	5	7

Rank in order 1 to 12 where 1=Best and 12=Worst.
Use each Forced Ranking (1, 2, 3, … 12) ONLY ONCE (No ties)

Sample	Rank
1	_____
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____
11	_____
12	_____

Figure 3 - Sample Judging Sheet

The Choice Design was created in JMP. A fixed Random Seed = 12 was used.

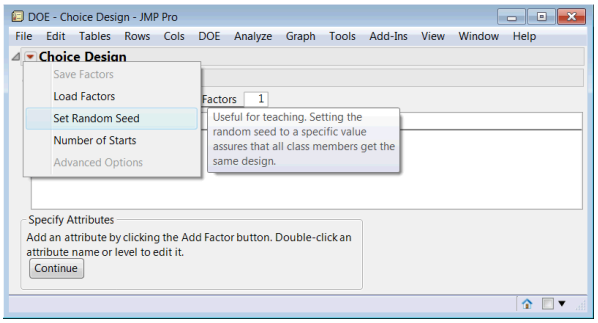


Figure 4 - Set Random Seed

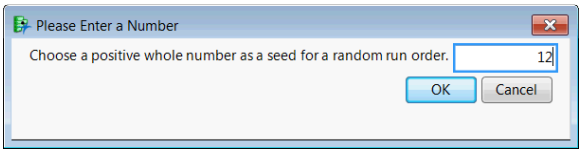


Figure 5 - Random Seed Entry

The Attribute was a 12 Level Factor, which was simply the sample number, 1-12.

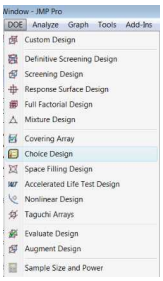


Figure 6 - Choice Design option under DOE

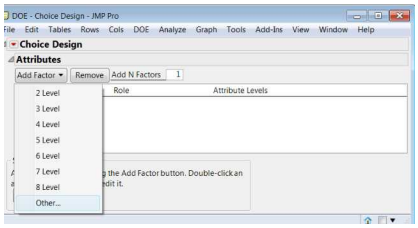


Figure 7 - Adding a Factor

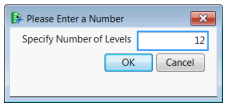


Figure 8 - Specifying 12 Levels

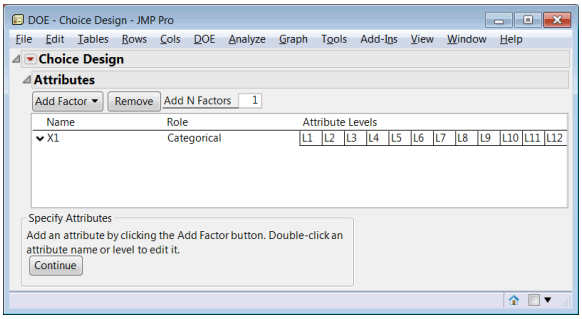


Figure 9 - Attribute and Levels

The DOE Model Controls were left as is, since there is only one term being modeled.

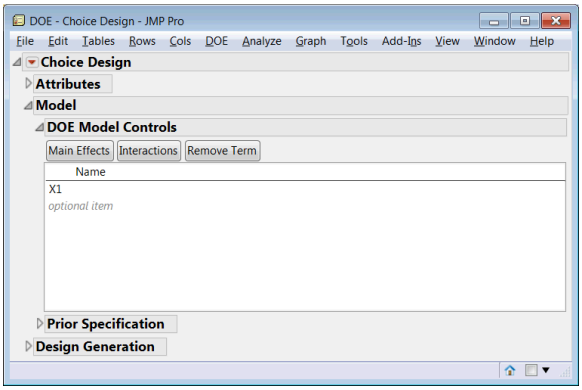


Figure 10 - DOE Model Controls

The Prior Specification was left at the default, since nothing was known about Prior Mean or

Prior Variance Matrix.

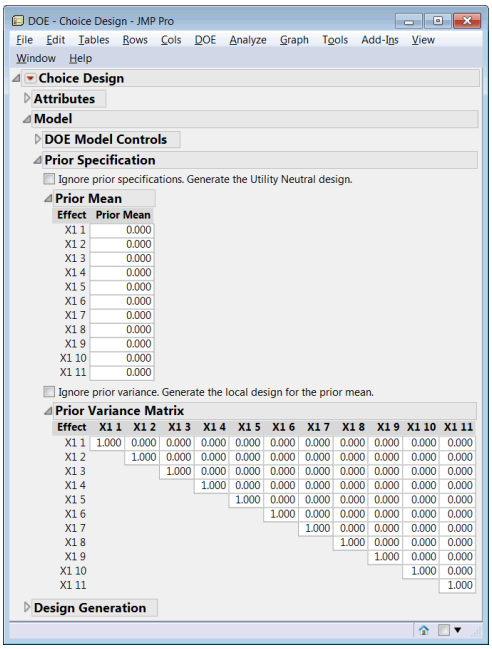


Figure 11 - Prior Specification

The Design Generation options are shown below. An explanation of the Design Generation settings follows.

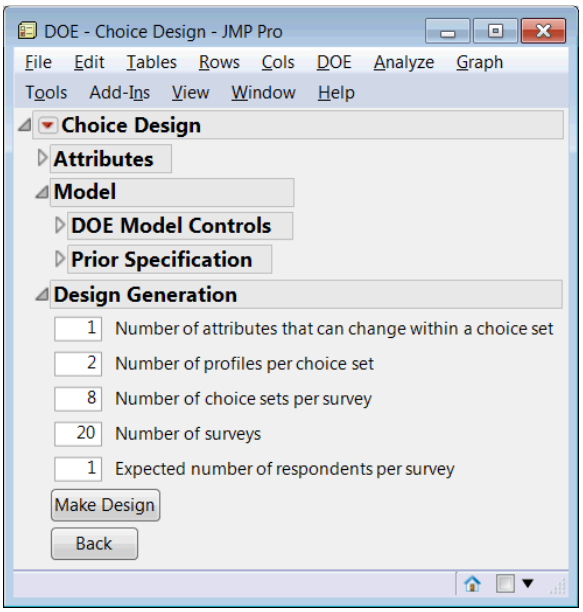


Figure 12 - Design Generation

- Number of attributes that can change within a choice set: 1 (Only the sample number will change.)
- Number of profiles per choice set: 2 (Two samples will be compared at a time.)
- Number of choice sets per survey: 8 (This seemed like a reasonable number of pairs to ask a judge to select a preference from. It was also the default.)
- Number of surveys: 20 (This provided 20 unique sets of pairings in anticipation of having 20 judges. There were 18 judges; sheets 7 and 15 went unused.)
- Expected number of respondents per survey: 1 (Each judge had a unique scoring sheet.)

Clicking [Make Design] generates the following design. (Only the first two surveys are shown.)

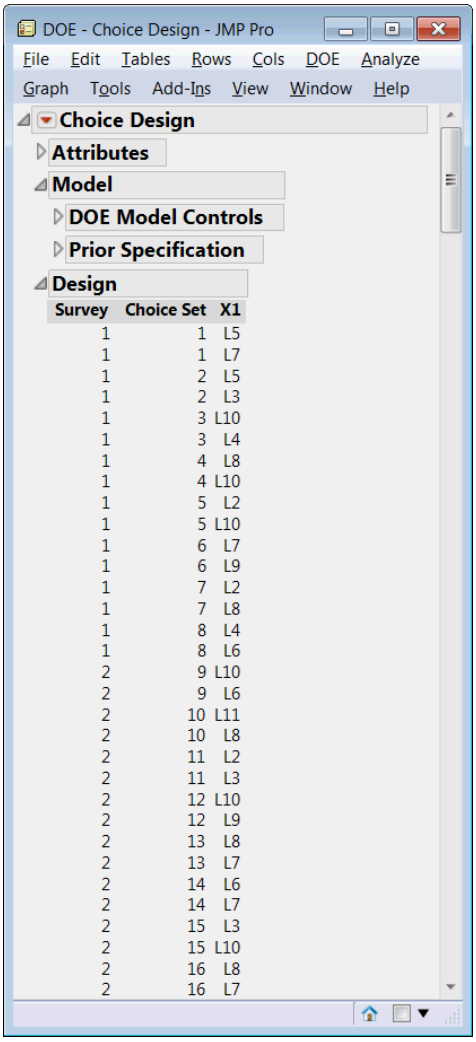


Figure 13 - Design

At the bottom of the Design is the [Make Table] button.

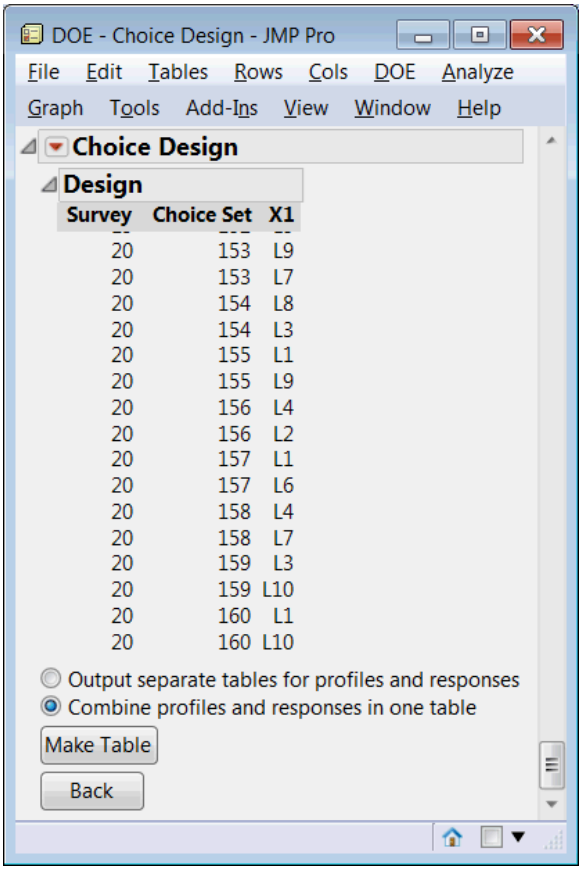


Figure 14 - Design Options

The following Table is generated. The Response Indicator will then be entered as 0 for the sample not chosen, and 1 for the sample chosen (in each pair).

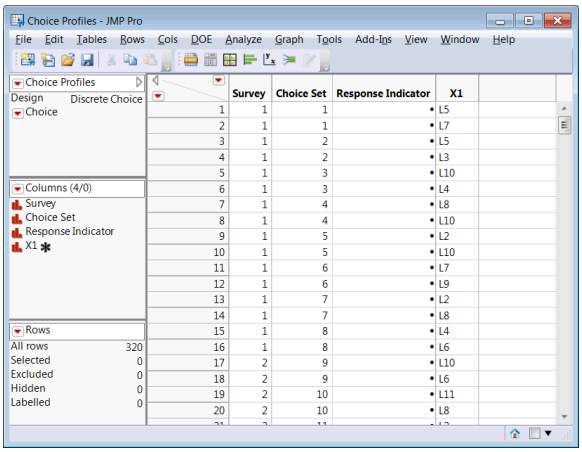


Figure 15 - Output from Make Table

Data from the 18 judges' scoring sheets were entered into JMP real-time as the scoring sheets were submitted. Judge sheets 7 and 15 were not used.

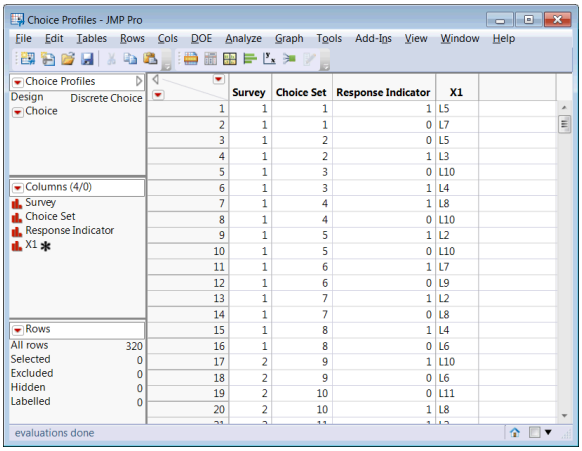


Figure 16 - Data Table

Once all data was entered, creating the model was simple. The Choice script is already in the data file. Simply run the script.

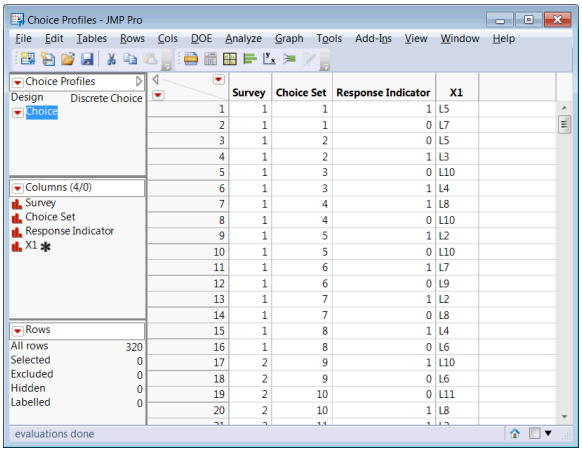


Figure 17 - Choice Script in Data Table

The Profile Data is already completed, based on how the Choice Design was set up.

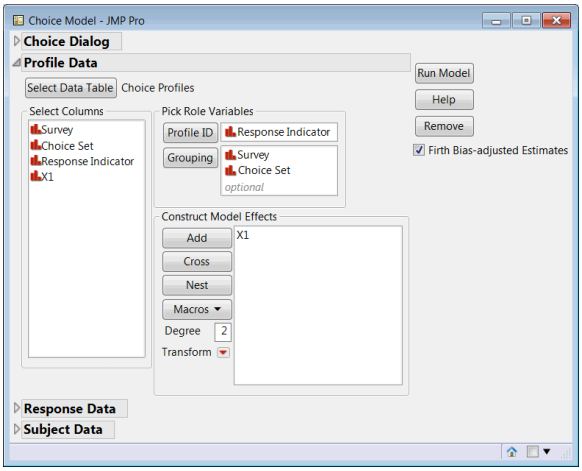


Figure 18 - Profile Data

JMP will ask you if all your data is in one table.

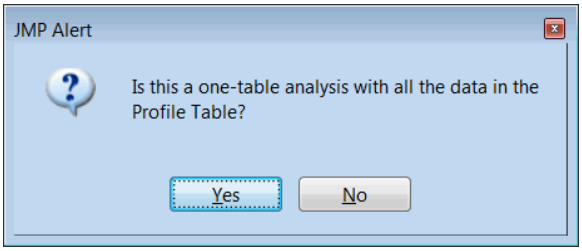


Figure 19 - JMP Alert

The Choice Model is generated, with the Estimate and Standard Error for each term. A high estimate is an indication that the sample was more preferred.

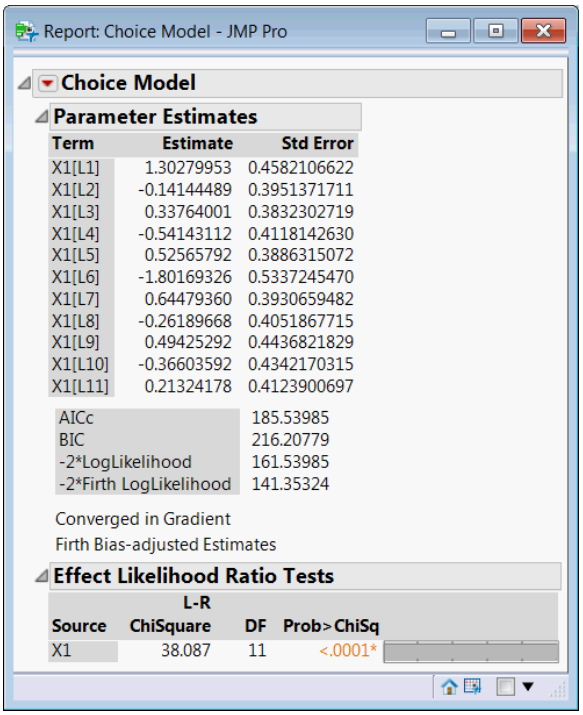


Figure 20 - Choice Model Parameter Estimates

The resulting parameter estimates are referred to as Utility, the level of satisfaction consumers receive from products with specific attributes. The Utility Profiler shows this information graphically.

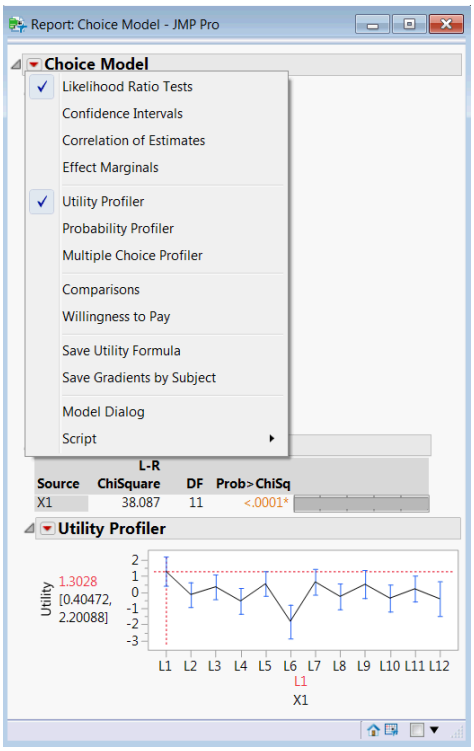


Figure 21 - Probability Profiler Menu Location

The Probability Profiler is another available profiler. Honestly, at first it was not understood what the Probability Profiler was showing. That's the beauty of JMP; it's easy to learn just by reading the help documentation that comes with JMP. With just a few minutes of reading, it was easy to determine that the Probability Profiler shows the probability of selecting one choice over the baseline (which can quickly be changed). For example, by setting the baseline to L7, it's simple to observe the probability of L1 being selected over L7 is 0.659.

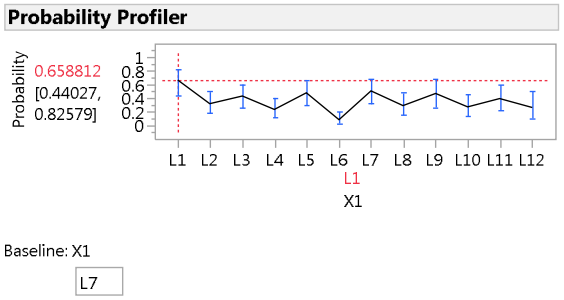


Figure 22 - Probability Profiler

Let's compare that to what actually happened. As George Box said, "All models are wrong. It's the data that are real. They actually happened!" (Box, Hunter, & Hunter, 2005) With 144 pairings and 66 possible paring combinations, we only expect about two parings of the same wines. It turns out wine #1 faced off against wine #7 twice, and beat it both times. This is insufficient data to conclude much, but it was interesting to check.

Forced Ranking

Forced ranking involves arranging n items in order of preference from 1 to n without any ties.

Forced rankings were entered into JMP as shown. Again, the data entry was conveniently performed real-time as sheets were turned in. (The data entry clerk was Don's son Nick, who has been using JMP since first grade.)

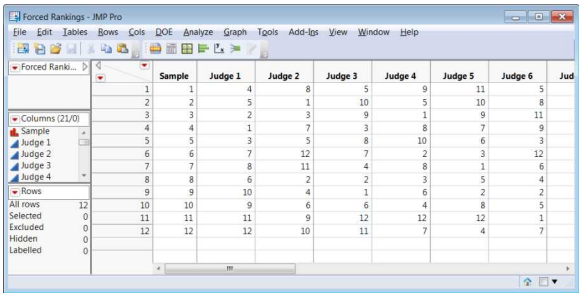


Figure 23 - Data Table

The data were stacked so that a Fit Y by X plot could be generated.

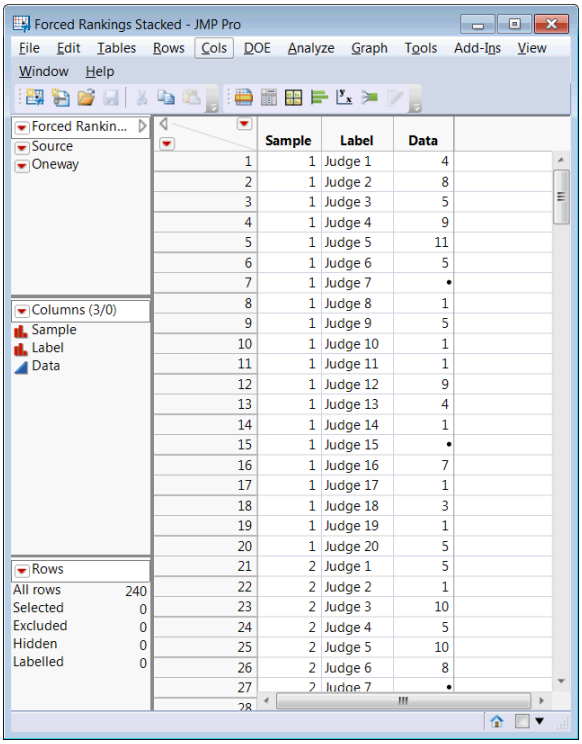


Figure 24 - Stacked Data

Recall that judging sheets 7 and 15 were not used, since 20 judging sheets were generated and only 18 judges were present.

A simple Fit Y by X reveals that sample #1 was ranked highest. (Note the scale was reversed to improve visualization of the rankings.)

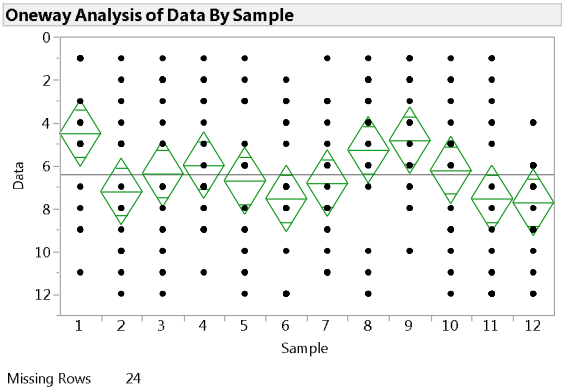


Figure 25 - Fit Y by X

This Oneway Analysis treats ordinal data as continuous. However, this seems reasonable since the attempt is to compare rankings rather than estimate means. Treating the data as an ordinal data type results in a Mosaic Plot that can at best be manually interpreted. (The colors were reversed to give positive meaning to reds and negative meaning to blues.)

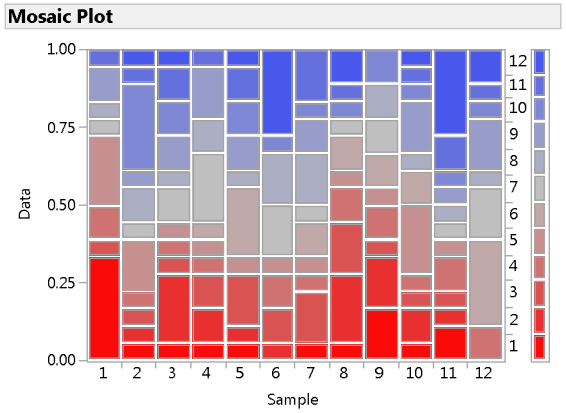


Figure 26 - Mosaic Plot

It is a bit of a challenge to accurately determine preferences from the Mosaic Plot, although it clearly visualizes some differences. For example, Samples #1, #8 and #9 appear to be more desirable. Samples #6 and #11 both have similar undesirable rankings of 12 (dark blue); looking at the remainder of the rankings for those samples, it appears sample #11 has more favorable rankings (red) but also seems to have more unfavorable rankings (blue). Treating the data as continuous makes it easier to justify calling these a tie, since the means are not statistically different. The Connecting Letters Report shows that sample #6 is only statistically different from samples #1, #8, and #9.

Connecting Letters Report

Level					Mean
12	A				7.7222222
6	A				7.5555556
11	A				7.5555556
2	A	B			7.2222222
7	A	B	C		6.8333333
5	A	B	C		6.7222222
3	A	B	C	D	6.3888889
10	A	B	C	D	6.2222222
4	A	B	C	D	6.0000000
8		B	C	D	5.2777778
9			C	D	4.8333333
1				D	4.5000000

Levels not connected by same letter are significantly different.

Figure 27 - Connecting Letters Report

The Connecting Letters Report also shows that although Sample #1 was rated the highest, it is not statistically different from samples #3, #4, #8, #9, or #10.

Issues with Forced Ranking

Forced Ranking seems straightforward, but many questions were asked by judges during the event; this indicated that a quality check of the data might be needed. It's easy to check which judges followed directions. The simplest check is to visually look at the data.

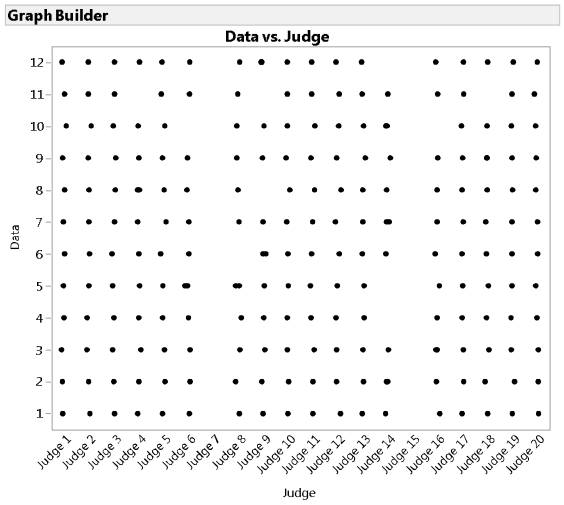


Figure 28 - Data vs. Judge

Recall that Judge 7 and Judge 15 sheets were not used. It's clear that many judges deviated from the directions; many used the same ranking twice and excluded some rankings. In fact, 7 of the 18 judges did not correctly use the rankings of 1-12 as instructed.

Judge 14 had the most issues with the Forced Ranking, using three of the ranks twice and omitting three other ranks. This indicates that Judge 14 did the Forced Ranking after completing the Discrete Choice portion of the judging, for obvious reasons.

Simple Rating Scales

Issues with Simple Rating Scales

Unconstrained rating scales, such as 0-10, have issues as well. You may be thinking, "Why not just use a 0-10 scale, allowing the judge to assign any value to any sample?" A previous event revealed issues with this methodology. One issue was inconsistency in the use of this scale. Judges were instructed, *"We encourage you to use the entire scale! It helps with accurately ranking them. By definition, less than half of them should be above average … right?!"*

The following rating guide was also provided, clearly emphasizing that 5 is average:

0	Poured it out
1	I'll drink it if I'm thirsty
2	OK, but not something I'd buy
3	Not quite as good as most of the others
4	Almost average
5	Average
6	Maybe a little better than average
7	Above average
8	Much better than the rest
9	Excellent, but not perfect
10	I've never tasted anything this good!

Even with such clear directions striving for a distribution centered at 5 and symmetrically spread across the scale, the judges provided the following data.

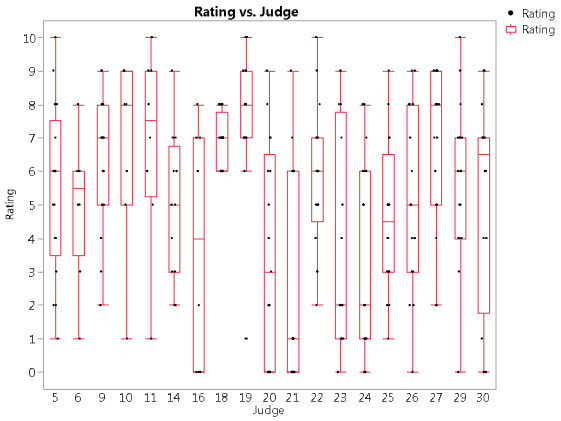


Figure 29 - Rating vs. Judge with Box Plots

Judges 18 and 19 ranked all but 1 sample above average, even though the judging sheet clearly stated, *"… less than half … should be above average …"*!

Many judges also did not complete the judging sheet. Only 10 of 19 judges scored all 17 samples. The number of samples scored is shown below.

Judge	N
5	17
6	8
9	17
10	7
11	8
14	12
16	8
18	16
19	17
20	17
21	17
22	17
23	16
24	17
25	16
26	17
27	17
29	17
30	16

Figure 30 - Quantity of Samples Scored, by Judge

The ratings were bimodally distributed, with one distribution centered at 1.2 and the other centered at 6.6.

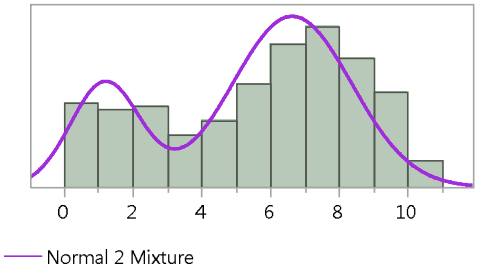


Figure 31 - Bimodal Distribution

There was a tendency to overrate samples; 55% of the ratings were above average (151/277) while only 35% were below average (98/277).

It has also been shown that judges tend to avoid extreme ratings in the beginning of a series of evaluations, and expand their ratings as they proceed. (Unkelbach, Ostheimer, Fasold, & Memmert, 2012) This could not be tested with our data, since judges were not instructed to rate the samples in any particular order.

The judging method in this previous event did not result in a drastic separation of the good from the bad. In fact, the top rated sample was only statistically different from one other sample; likewise, the lowest rated sample was only statistically different from one other sample. The only two samples that were significantly different were samples #3 and #11.

Connecting Letters Report

Level		Mean
11	A	6.3529412
17	A B	6.2142857
15	A B	6.1176471
7	A B	6.0000000
4	A B	5.8125000
10	A B	5.6875000
6	A B	5.5555556
2	A B	5.5000000
12	A B	5.3333333
13	A B	5.2941176
8	A B	5.2222222
14	A B	5.1764706
9	A B	4.7857143
5	A B	4.6000000
1	A B	4.2666667
16	A B	3.8235294
3	B	2.9411765

Levels not connected by same letter are significantly different.

Figure 32 - Connecting Letters Report

In this previous experiment, samples #5 and #16 were the same wine. It is entertaining to look how the judges scored the same wine. The ratings for samples #5 and #16, by Judge, are shown below. (The authors take pride in being consistent, scoring them either the same or a difference of 1.)

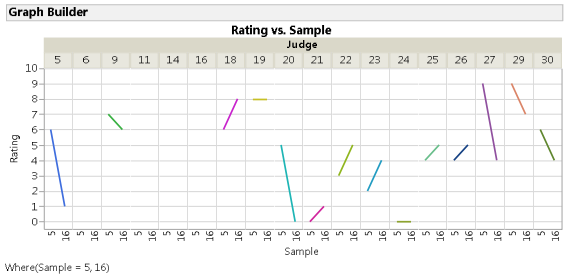


Figure 33- Ratings of Same Two Wines, by Judge

A more reliable model could be produced by transforming ratings data into inferred rankings. This would require much less demanding assumptions about the properties of the responses. (Louviere, Hensher, & Swait, 2000)

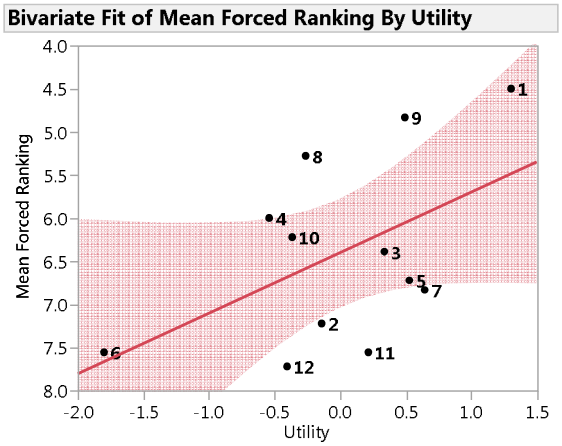
Comparison of Results from Discrete Choice and Forced Ranking Methodologies

The Discrete Choice methodology provided a measure of Utility, ranging from -1.80 to 1.30. The Forced Ranking methodology yielded Mean Ranking ranging from 4.50 to 7.72. Although the scales and units are different, a comparison can still be made. The tabulated results are as shown.

Sample	Mean Forced Ranking	Utility
1	4.500	1.303
2	7.222	-0.141
3	6.389	0.338
4	6.000	-0.541
5	6.722	0.526
6	7.556	-1.802
7	6.833	0.645
8	5.278	-0.262
9	4.833	0.494
10	6.222	-0.366
11	7.556	0.213
12	7.722	-0.406

Figure 34 - Mean Forced Ranking and Utility

Graphically, the correlation is shown below.



Linear Fit

Mean Forced Ranking = 6.4027778 - 0.7020522*Utility

Summary of Fit

RSquare	0.256703
RSquare Adj	0.182373
Root Mean Square Error	0.979287
Mean of Response	6.402778
Observations (or Sum Wgts)	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	3.311981	3.31198	3.4536
Error	10	9.590025	0.95900	Prob > F
C. Total	11	12.902006		0.0928

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6.4027778	0.282696	22.65	<.0001*
Utility	-0.702052	0.377777	-1.86	0.0928

Figure 35 - Correlation of Mean Forced Ranking and Utility

The correlation is weak, with RSquare Adj = 0.18. The correlation is not statistically significant, at 95% confidence. (p=0.0928.)

For the purpose of visual clarity, the Mean Forced Ranking and Utility values can be normalized from 0 to 1, but the plot and statistics remain the same.

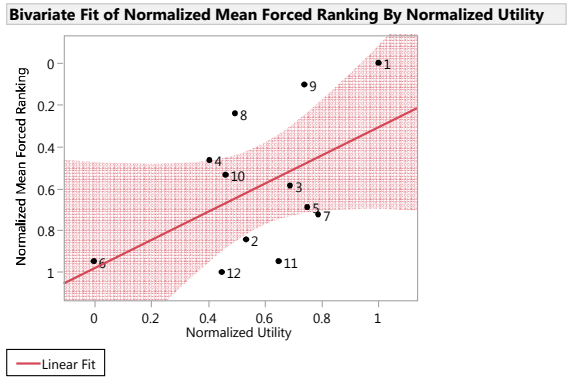


Figure 36 - Normalized Scales

Why was there so little correlation between the Discrete Choice and Forced Ranking methodologies? Perhaps judges were inconsistent in their preferences. To check that theory, we can compare their Discrete Choice results to what the Discrete Choice results would have been if they were derived from the Forced Rankings supplied by the judges.

The results of that comparison are shown below. Only 5 of the judges' Discrete Choices agreed with their Forced Ranking for all samples. Judge 16 had a surprising 7 out of 8 disagreements; their paired preference was different from their forced ranking in 7 out of 8 trials!

Tabulate

Survey	Disagreement Sum
1	0
2	0
3	5
4	4
5	2
6	0
7	.
8	1
9	4
10	0
11	0
12	6
13	2
14	5
15	.
16	7
17	1
18	2
19	2
20	4

Figure 37 - Number of Choices that Disagreed with Forced Ranking (out of eight), by Judge

There were many surprising extremes, such as Judge 9's ranking of samples #6 and #10 vs. their preference when paired. Judge 9 ranked sample #6 as 12th, and ranked sample #10 as 1st, yet preferred sample #6 when individually paired against sample #10. Perhaps the two samples are so similar that the same judge would have a different preference each time. Perhaps, and more likely, the judge was simply doing too much sampling and not enough judging!

The resulting Discrete Choice model when using the response derived from the Forced Ranking is shown below.

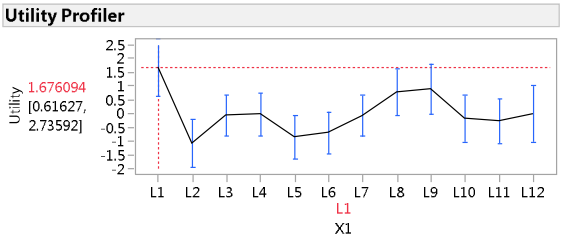


Figure 38 - Utility Profiler – Derived Choice Data

Recall that the original Discrete Choice model yielded the result below, for comparison.

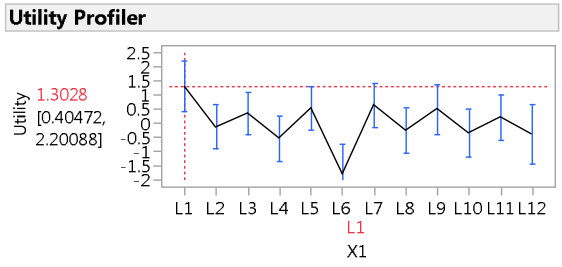


Figure 39 - Utility Profiler - Original Choice Data

The differences are somewhat surprising. The inconsistencies within each judge may be the cause of the lack of strong correlation between the two methods. To investigate this further, let's look at the Utility derived from the Forced Ranking and compare it to the Mean Forced Ranking.

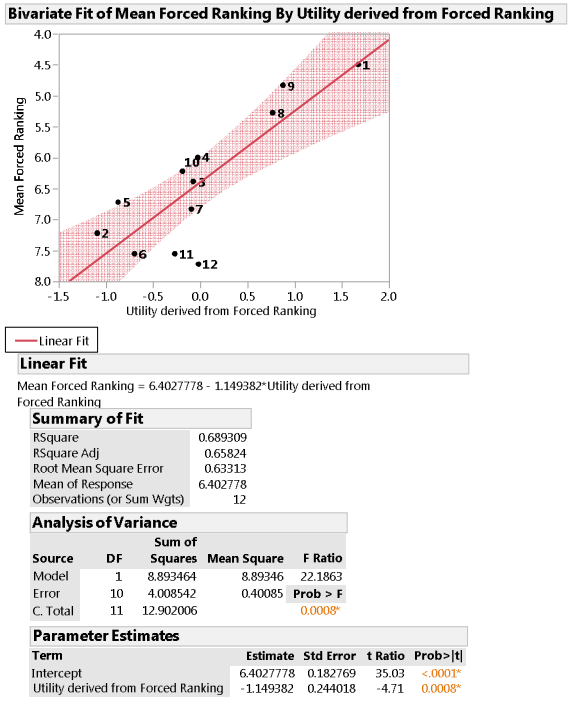


Figure 40 - Correlation of Forced Ranking to Utility Derived from that Ranking

The models seem to agree quite well! This is the result we would have seen if the judges had agreed with themselves when doing both the Forced ranking and Discrete Choice scoring. The RSsquare Adj is 0.658 (compared to 0.182 seen in the original data). It seems that the disagreement between the Discrete Choice results and the Forced Ranking results were indeed caused by the inconsistencies within each judge's own results! **Recall that only 5 of the 18 judges had all 8 pairs in the Discrete Choice methodology agree with their Forced Ranking.**

Inherent Issues with the Methodology

As with most experiments, there were some issues with the methodology worth identifying. Although the samples were wrapped in foil and labeled, the bottle shapes were not completely concealed. Most of the bottles were similar in shape, although not identical. It is possible that a wine connoisseur (of which there were many) could identify a wine by the shape of its bottle, thus biasing the data for that judge.

There is an inherent bias towards sweeter drinks when only sips are involved. Recall that Pepsi claims to have won the challenge, even though Coke continuously outsells Pepsi and beats Pepsi in Home Use Tests (HUT). (Gladwell, 2005)

Judges were free to wander, and they naturally compared notes and offered tips to each other. "You really need to try #2!" The assumption of independent observations is a stretch. A party where judges were not allowed to communicate would not be much fun, and would likely be the last of such parties that these judges would attend!

This experiment was a Central Location Test (CLT). Results are more realistic with a HUT. Consumers are known to have much different preferences when allowed the time to consume an entire glass of each sample in their own home before evaluating it. (Boutrolle, Delarue, Arranz, Rogeaux, & Koster, 2007)

Conclusion

A comparison was made between two different judging methodologies, a) the DOE Choice Design feature in JMP and b) a Forced Ranking methodology, for determining order of preference of taste. Each of the two methodologies provided an order of ranking. The DOE Choice Design provided a measure of Utility for each sample. The Forced Ranking provided a Mean Ranking for each sample (the average of all the forced rankings for the sample).

The results of the two methodologies were compared and found to be quite different, with very little correlation. An in-depth look discovered that the primary cause of the disagreement between the results of the two methodologies was due to judging inconsistencies. The judges were not consistent with their preferences between the two judging methodologies; many of the judges did not even agree with themselves! Only 5 out of 18 judges had discrete choice results that completely agreed with their forced rankings.

When the Forced Rankings were used to create a *new* "derived" Discrete Choice scoring for each judge, the agreement between the two models greatly improved. The correlation between the Utility in the "derived" Discrete Choice methodology and the Mean Ranking from the Forced Ranking methodology improved the RSquare Adj from 0.18 in the original scoring to 0.69 in the "derived" Forced Ranking.

Which methodology provides the most useful model? Since it is known that a Choice Design is typically easier for judges to complete (Meyer, 2012), the DOE Choice Design seems to be the better model. Intuitively, judges are more likely to identify the best of two samples than they are to correctly order 12 samples.

The DOE Choice Design should be considered in situations where Forced Ranking may prove to be difficult. It was easier for judges to complete in this case. Task difficulty increases considerably with the number of samples to be ranked. Reliability of Forced Ranking decreases with more options. (Louviere, Hensher, & Swait, 2000)

A rating judging methodology was also analyzed to show the inadequacies of simply allowing each sample to be freely rated from 0 to 10. Such ratings also come with demanding assumptions that humans must meet while generating rating data; meeting these assumptions is not necessary with Forced Rankings or Discrete Choices. (Louviere, Hensher, & Swait, 2000)

Bibliography

Boutrolle, I., Delarue, J., Arranz, D., Rogeaux, M., & Koster, E. P. (2007, April). Central location test vs. home use test: Contrasting results depending on product type. *Food Quality and Preference, 18*(3), 490-499.

Box, G. E., Hunter, J. S., & Hunter, W. G. (2005). Statistics for Experimenters. Hoboken, NJ.

Gladwell, M. (2005). *Blink.* New York, NY: Back Bay Books.

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated Choice Methods.* New York: Cambridge University Press.

Meyer, C. D. (2012). *Who's #1? The Science of Rating and Ranking.* Princeton University Press.

Unkelbach, C., Ostheimer, V., Fasold, F., & Memmert, D. (2012, September). A calibration explanation of serial position effects in evaluative judgments. *Organizational Behavior and Human Decision Processes, 119*(1), 103-113.