Sawtooth Software

RESEARCH PAPER SERIES

Anchoring Maximum Difference Scaling Against a Threshold – Dual Response and Direct Binary Responses

Kevin Lattery,
Maritz Research

530 W. Fir St.
Sequim, WA 98382
(360) 681-2300
www.sawtoothsoftware.com

[1]ANCHORING MAXIMUM DIFFERENCE SCALING AGAINST A THRESHOLD – DUAL RESPONSE AND DIRECT BINARY RESPONSES

KEVIN LATTERY
MARITZ RESEARCH

I. INTRODUCTION

Maximum Difference Scaling is a choice based tradeoff technique for understanding the relative value of several attributes. Respondents are asked to choose the "best" and "worst" attribute from a subset of the attributes. An example of a MaxDiff question is this:

Thinking of your ideal Mobile Phone Retail Store, which of these features is most important and which is least important to you?

	Most Important	Least Important
Friendly sales representative	◉	○
Store front is attractive	○	◉
Convenient store location	○	○
Information about rebates/discounts	○	○

Respondents see several screens like this, each time choosing their best and worst attribute. This kind of technique is useful because it does not depend upon how respondents use a scale. Instead it asks respondents to make choices. So it is in the family of tradeoff techniques and shares similarities with conjoint analysis. Like a conjoint, MaxDiff also has an experimental design and can be analyzed using the same techniques as conjoint. Indeed the most common form of analysis is Hierarchical Bayes (HB) such as Sawtooth's CBC/HB module (which was what we used in this paper).

What we understand from MaxDiff is the relative value of each of the attributes. To make this point clearer, consider the following thought experiment. Imagine two respondents, call them Brad and Angelina, who would rank order the attributes the same way. They would therefore answer each MaxDiff task the same way, and as a result we would derive the same utilities for both of them (within error). But as it turns out, Brad and Angelina are very different. For Angelina, all of the attributes are important, while for Brad none of them matter. So while the rank order is the same, all of Brad's utilities should be shifted lower, in fact much lower than Angelina's utilities.

In some cases, relative utilities are fine, but sometimes researchers want the utilities to take into account some kind of absolute measure. That is, one may want Brad and Angelina's utilities

[1] Originally published in the 2010 Sawtooth Software Conference Proceedings.

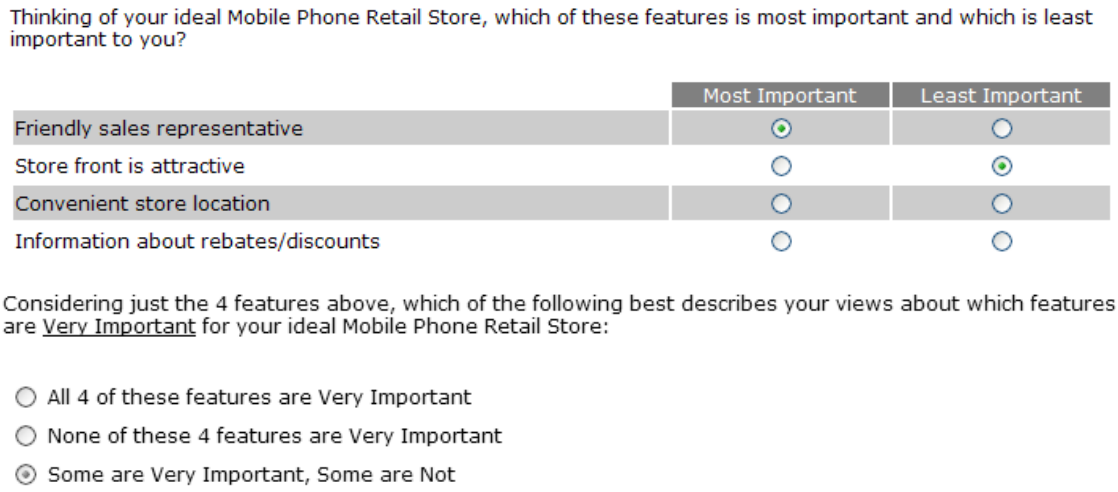
to be different because Brad really thinks none of the attributes are important, while Angelina does.

One method to make MaxDiff utilities less relative is to anchor the utilities to a specific point. For instance, we might make a utility of 0 to be a reference point, where above 0 means important and below 0 means unimportant. This means all of Brad's utilities would be shifted below 0 while Angelina's would all be above 0.

This is known as anchoring the utilities to a threshold. In the example above 0 was a threshold to which the utilities were anchored. In the remainder of this paper, we will be discussing two methods for anchoring utilities to a threshold of 0: Indirect Dual Response and Direct Binary Responses.

II. Anchoring Technique One: Indirect Dual Response Method

This method was first suggested by Jordan Louviere. After each MaxDiff task, one asks a follow up question about whether all, none, or some of the attributes meet a threshold. An example of this follow up question is shown below:



In this case the follow up question asks whether the attributes are "Very Important", but any other phrase could be used. This will become the anchoring that corresponds with a utility of 0. So in this case attributes with a utility above 0 are "Very Important", while attributes with a negative utility are not "Very Important".

Implementing this method requires some clever coding. First, one no longer uses a reference level. For the best and worst pick, one uses full dummy coding. The example below will show how a specific task is coded. For this example, assume there are 8 attributes, and the respondent saw attributes 1, 3, 6, and 8.

For the best choice we have the following coding, which is the same as typical MaxDiff coding without a reference level:

a1	a2	a3	a4	a5	a6	a7	a8
1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1

For the worst choice we also have the typical MaxDiff coding but without a reference level:

a1	a2	a3	a4	a5	a6	a7	a8
-1	0	0	0	0	0	0	0
0	0	-1	0	0	0	0	0
0	0	0	0	0	-1	0	0
0	0	0	0	0	0	0	-1

The trickier part is how to code the follow up question.

If the respondent said “None are Very Important” then one added the following task:

a1	a2	a3	a4	a5	a6	a7	a8
1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0

We pretend the respondent saw 5 attributes, with the 5th fictional attribute winning. The idea here is that each of the attributes loses to the zero vector and therefore the utilities will be negative.

If the respondent said “All are Very Important” then we add the following task instead:

a1	a2	a3	a4	a5	a6	a7	a8
-1	0	0	0	0	0	0	0
0	0	-1	0	0	0	0	0
0	0	0	0	0	-1	0	0
0	0	0	0	0	0	0	-1
0	0	0	0	0	0	0	0

Again we pretend the respondent saw 5 attributes with the 5th fictional attribute winning. In this case, the negated attributes lose to the zero vector, meaning they are positive.

The initial coding suggested by Sawtooth Software added no additional information when the respondent said "Some are Very Important, Some are Not". While developing the presentation for the 2010 Sawtooth conference this coding was seen as incomplete. Later in this paper we will show why this incomplete coding should not be used.

The more complete coding was suggested by Paul Johnson of Western Watts. This modifies the initial coding of the Best and Worst tasks. Using the same example, we would alter the initial Best task to the following:

a1	a2	a3	a4	a5	a6	a7	a8
1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0

This is the same as the "None are Very Important", except that the winner will be the actual best attribute (rather than the zero vector). The idea here is that we know some of the attributes are very important, which means that the attribute selected beats the zero vector. That is the additional information added here.

We also need to modify the Worst task in the same way:

a1	a2	a3	a4	a5	a6	a7	a8
-1	0	0	0	0	0	0	0
0	0	-1	0	0	0	0	0
0	0	0	0	0	-1	0	0
0	0	0	0	0	0	0	-1
0	0	0	0	0	0	0	0

The winner will be the actual worst attribute. This means the negative of the worst attribute beats the zero vector – suggesting the worst attribute is less than zero.

In summary the revised coding tells us that the best attribute beats the zero vector, while the worst attribute loses to the zero vector. The other attributes we still know nothing about whether they are positive or negative. This additional information is imperative to properly anchor the MaxDiff utilities. While the revised coding provides much more information, it should be noted that we may not gather threshold information about some attributes. If each time an attribute appears it is neither best nor worst, and if the follow up is "Some are, Some are not", then we know nothing about whether the attribute is positive or negative.

III. Anchoring Technique Two: Direct Method

The indirect method requires a follow up question with each MaxDiff task. In addition we also may not gather information about whether some of the attributes are positive or negative. This leads us to consider another technique, which simply asks the respondent to check whether each attribute is above or below the threshold. An example of this direct method is:

Please tell us which of the features below are Very Important for your ideal Mobile Phone Retail Store?
(Check all that are Very Important)

- ☐ Good layout and design of the store
- ☐ Do not have to wait for service
- ☑ Variety of accessories available
- ☐ Store front is attractive
- ☐ Store has the phones that I want
- ☑ Store has the carrier(s) (AT&T, Verizon, etc) I want
- ☑ Clarity of displays and product information
- ☐ Accessibility of phones for you to try
- ☐ Informational materials available
- ☐ Convenient store hours (evenings/weekends, etc.)

- ☐ I do not consider any of these to be Very Important

This question may be asked after all the MaxDiff tasks. This means no break in the continuity of MaxDiff tasks, and less time than the indirect dual response method. Perhaps most importantly, we get information about whether each attribute is above or below a threshold.

The coding used in this paper involved adding two tasks for each respondent: one representing the attributes above the threshold, and one for the attributes below the threshold. To illustrate this coding, assume there are 8 attributes, and that attributes 1, 3, 6, 7, and 8 meet the threshold of "Very Important". Then we add the following task:

a1	a2	a3	a4	a5	a6	a7	a8
-1	0	0	0	0	0	0	0
0	0	-1	0	0	0	0	0
0	0	0	0	0	-1	0	0
0	0	0	0	0	0	-1	0
0	0	0	0	0	0	0	-1
0	0	0	0	0	0	0	0

The zero vector (last row) wins, meaning that the negations of the utilities lose to zero. Again we have no reference level. The remaining attributes do not meet the threshold and are coded with positive ones losing to the zero vector:

a1	a2	a3	a4	a5	a6	a7	a8
0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0

Adding these simple two tasks informs the model whether each attribute should be positive (meets threshold) or negative (does not meet threshold). Of course if all of the attributes lie on the same side of the threshold then only one task would be added.

Alternative codings were tested, including the binary version where each attribute was compared with a zero vector. This resulted in slightly different utilities, primarily increasing their variance.

IV. Results – Timing and Satisfaction

563 respondents did the direct method only, while 569 respondents did the indirect dual response augment after each MaxDiff task, followed by the direct augment after all MaxDiff tasks.

Comparing these two groups, the direct method is much quicker:

1. Respondents took an average of 4.3 seconds per task to complete the indirect dual response question. So with 15 tasks, the total time is 1 minute 21 seconds. This time computation includes removing 10% of outlier respondents who took more than 40 seconds per task.
2. In comparison the 20 attribute grid with 10 per screen took about 19 seconds of total time.

Given the additional time of the indirect augment, coupled with the dual response break in continuity, we expected respondents to be less satisfied with the survey when they were asked the indirect dual response augment. However, we did not observe any significant change in satisfaction with the survey. On a typical five-point satisfaction scale, the Direct Method shows a slightly higher mean satisfaction of 4.08 vs. 4.00, and a 76% top 2 box score vs. 74% for the Indirect augment.

Overview of Sampling Split - 4 Primary Cells

To compare the two methods most directly, we will focus on the 569 respondents who completed both the indirect dual augment and the direct method. These respondents were assigned to one of the following four cells:

Group	N Size	Attributes	Items Per Task	MaxDiff Tasks
1	163	All 20	4	15
2	129	All 20	5	12
3	142	Better 12	4	9
4	135	Worst 12	4	9

Group 1 will be used for initial comparison and is our baseline. Group 2 is like group 1, but 5 attributes were shown at a time. Group 3 and 4 split will be compared with group 1 to see how well a subset of attributes matches the entire attribute list (more on this later).

We also showed 563 respondents the direct method only. This was done to see if the indirect dual response augment had any measureable impact on the direct results. It did not. So in order to compare the methods in the most direct fashion we will focus on these 4 cells above where respondents completed both methods.

V. Results – Group 1 Baseline

A. Observed Patterns of Choices

Among the 163 respondents of Group 1, we observed the following general patterns:

- 17% of respondents always choose a Mix (some important/some not)
- 72% of respondents use "All Very Important" at least once
    - 61% use at least twice
    - 48% use at least thrice
- 30% of respondents use "None Very Important" at least once
    - 17% use at least twice
    - 13% use at least thrice

So respondents are clearly using the different options in the dual response, sometimes choosing a mix, and other times an "All" or "None". But if we consider all of the tasks across all of the respondents we get the following breakdown of clicks:

All Very Important	**22.2%**
None Very Important	**5.9%**
Mix	**71.8%**

One can see that a Mix is clearly the most common click. This is in line with theoretical expectations. Showing four attributes at a time we should expect the Mix response about 30% to 90% of the time, depending upon how many attributes are above and below the threshold. The more evenly the attributes are distributed, the more Mix responses we expect, as the table below shows.

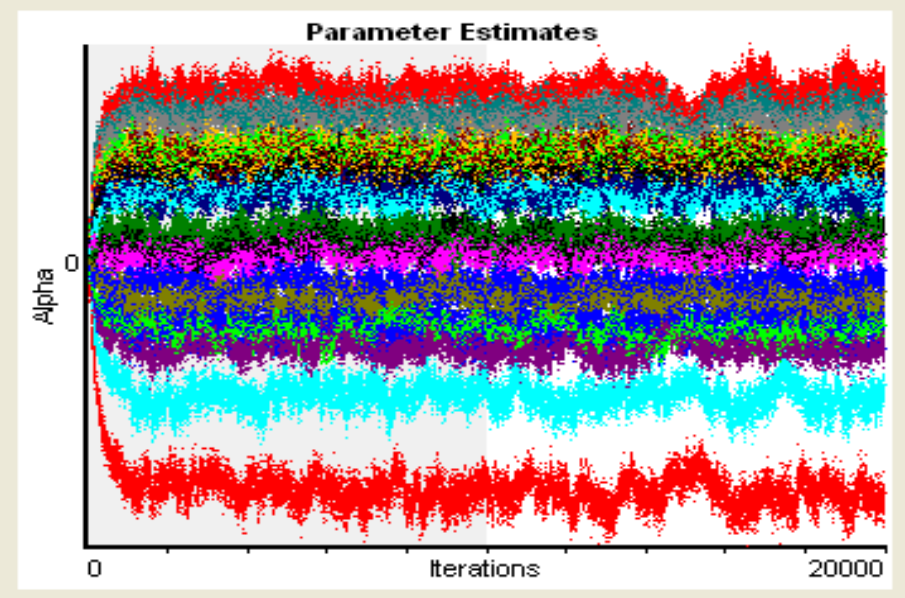
	Show 4 Attributes at a Time			**Show 5 Attributes at a Time**		
Percent Attributes Meeting Threshold	**Prob All > Threshold**	**Prob None> Threshold**	**Prob Mix**	**Prob All > Threshold**	**Prob None> Threshold**	**Prob Mix**
10%	0.0%	65.6%	**34.4%**	0.0%	59.0%	**41.0%**
20%	0.2%	41.0%	**58.9%**	0.0%	32.8%	**67.2%**
30%	0.8%	24.0%	**75.2%**	0.2%	16.8%	**83.0%**
40%	2.6%	13.0%	**84.5%**	1.0%	7.8%	**91.2%**
50%	6.3%	6.3%	**87.5%**	3.1%	3.1%	**93.8%**
60%	13.0%	2.6%	**84.5%**	7.8%	1.0%	**91.2%**
70%	24.0%	0.8%	**75.2%**	16.8%	0.2%	**83.0%**
80%	41.0%	0.2%	**58.9%**	32.8%	0.0%	**67.2%**
90%	65.6%	0.0%	**34.4%**	59.0%	0.0%	**41.0%**

In our case study, Group 2 with 5 attributes showed more Mix responses (79%), again as one would expect. Given the prevalence of Mix responses, it is clearly very important how one codes this information.

B. Convergence in HB

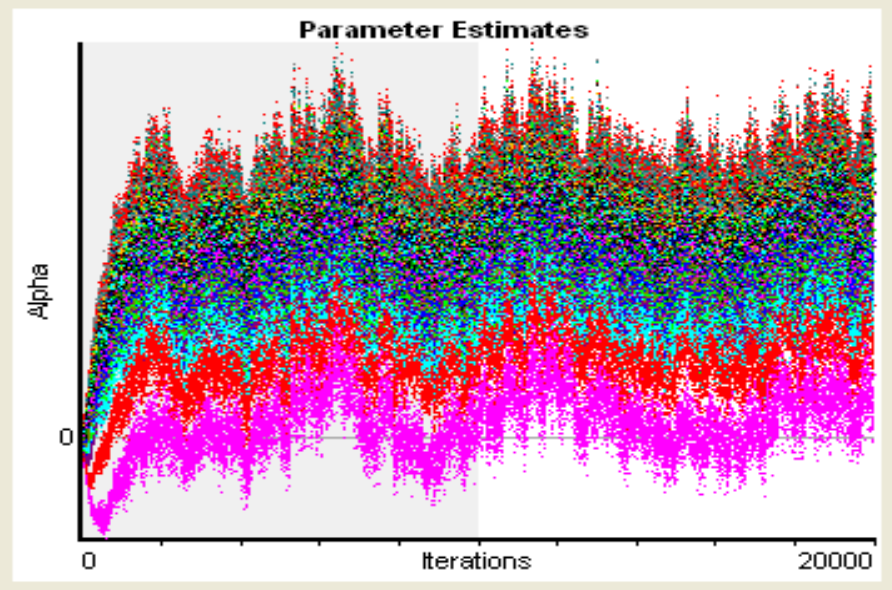
We estimated the utilities using Sawtooth Software's HB CBC, with a prior variance of 1. We first ran the normal MaxDiff utilities without any of the anchoring information. The utilities converged very nicely.

Only MaxDiff Questions



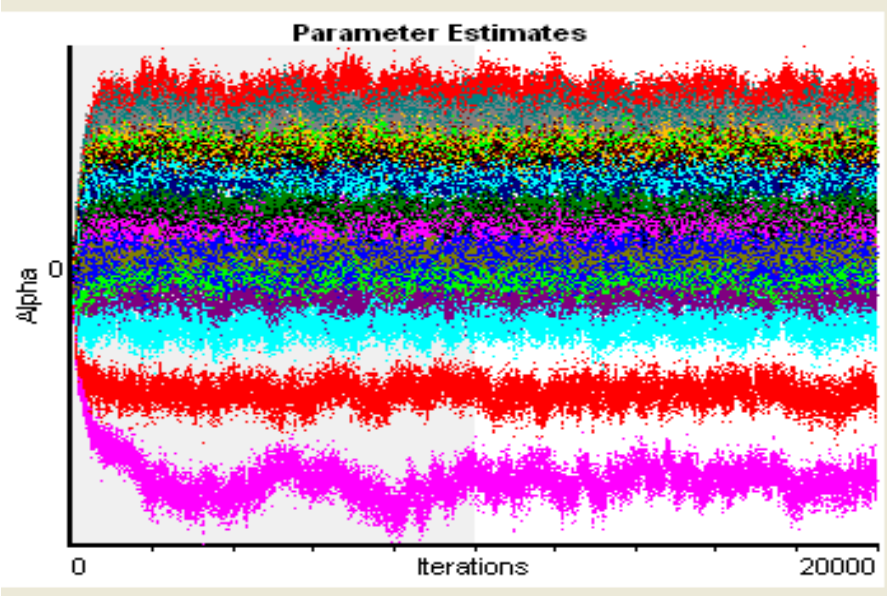
We then added the dual response augment. First we looked at the incomplete coding, where the "mix" response is not coded at all.

Indirect Dual Response Added – Incomplete Coding



As one can see, this did not converge. Playing with the degrees of freedom, prior variance, and number of iterations did not help with convergence. In comparison, when we coded the mixed responses using the revised coding of best and worst tasks, we once again got very nice convergence:

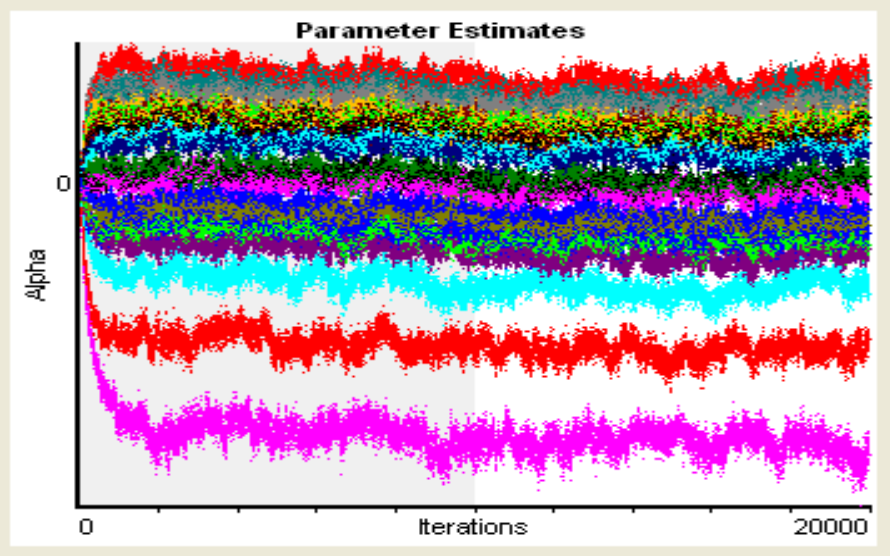
Indirect Dual Response Added – Complete Coding



This alone gave us good reason to implement the coding of the mixed response over no coding of the response. As we will see later, the incomplete coding really should not be used for many additional reasons as well.

Finally, we checked the Direct method where we asked respondents to check the attributes that were "Very Important". This also converged very nicely:

Direct Method



C. Utility Comparison

At the respondent level, the relative utilities from all four methods are nearly identical. If you rank the utilities and run a correlation between the ranks (at the respondent level), one gets an average correlation of .988 with the simple MaxDiff. So all the methods are preserving rank order of utilities. So although we included two holdout tasks with rankings, there was no difference in the ability to predict the rankings.

While the relative utilities of the methods are nearly identical, the absolute utilities are very different. The most important point is that the incomplete coding of the indirect dual response was a complete failure. This again is where we added no information at all when the respondent chose a "Mixed" response. To show just how badly this method failed consider the following table:

Indirect Augment	N	True Expect	Match	Comment
Always Positive	10	All Positive	10	
Always Negative	0	All Negative	0	
Always Mix (No Information)	28	Some Positive	16	10 all +, 2 all neg
Positive and Mixed Only	77		0	All 77 Positive
Negative and Mixed Only	17	Some Negative	0	Al 17 Negative
Positive and Negative (Opt Mix)	31		24	6 all +, 1 all neg
Total	163		50	

If the respondent thinks all the attributes are Very Important then the respondent will always give the positive response in the dual response, stating that all the attributes are very important. We see this happens for 10 respondents (first row of table), and the HB utilities match – giving all positive utilities. On the flip side, respondents who think none of the attributes are Very Important would always give the negative response to the dual response. There are no respondents in this group (2nd row of table) and the HB utilities also reflect that. So far, so good.

But in any other scenario we expect there to be some utilities for a respondent which are positive and some that are negative, reflecting that some attributes are Very Important while others are not. But in fact we rarely see this at all. In these cases, the lack of coding for a mixed response gives no information, and the attributes tend to inherent the non-mixed response from the respondent or the group response. For the 77 respondents who gave a positive and mixed dual response, all 77 had all positive utilities. From the standpoint of information in the model this is consistent, because the model is only seeing a few tasks which are stated to be all positive. The other tasks with a mixed dual response contain no information, which is consistent with a lower positive utility.

In total only 50 out of 163, or 30.7% of respondents have the correct utility structure of all positive/all negative/ or a mix of positive and negative. So we are not getting the anchoring right for the vast majority of respondents.

When we add the coding for the mix response, the results improve dramatically:

Indirect Augment	N	True Expect	Match	Comment
Always Positive	10	All Positive	10	
Always Negative	0	All Negative	0	
Always Mix (No Information)	28	Some Positive	28	
Positive and Mixed Only	77		77	
Negative and Mixed Only	17	Some Negative	17	
Positive and Negative (Opt Mix)	31		31	1 All Negative
Total	163		162	

All but one respondent is consistent in their utility structure of all positive/all negative/ or mix of positive and negative. This one exception was due to respondent inconsistency, where the respondent gave the same attributes an "All Positive" and "All Negative" response.

The Direct method matched the sign structure for all but 3 respondents (160 out of 163). These 3 exceptions were due to inconsistency in the respondent's choices, where the respondent said an attribute was Very Important but it lost to another attribute that was not Very Important.

The clear conclusion is that the incomplete coding of the indirect method is highly inadequate in capturing the mix of positive and negative utilities, where the other methods are extremely successful.

D. Simulated Data Comparisons

Using simulated data, we can show the incomplete coding of the indirect dual response to perform miserably, and that the results get worse as the number of Mixed responses increase. At this point however, we will no longer discuss the incomplete coding as we believe our discussion is sufficient to show it is completely inadequate.

Simulated data also shows that the Direct method is better than the Indirect Dual Response (complete coding). The reason for this is that Indirect method, even with the complete coding may still be indeterminate for some attributes. To better understand this, consider that each attribute is seen a certain number of times per respondent (for example 3 times). Each of those times, the follow up response could be the mixed response. If the attribute is not chosen as best or worst in any of those 3 scenarios, then we have no information about that attribute. This indeterminacy of the attributes increases with the number of attributes shown per task, and as the attributes are more evenly distributed (50% of attributes are positive and 50% negative). For this reason, we do not recommend the indirect method when there are 6 or more attributes shown per MaxDiff task.

The Direct method works extremely well with simulated data, outperforming the Indirect method in almost every set of simulated data. The only case in which the Direct method performs more poorly than the Indirect is when the true utilities of a respondent have small differences relative to the error.

Conclusion here is that in theory the direct method works best. The question is whether real people respond to the indirect augment more accurately than a list of attributes.

VI. Results – Group 3 and 4 Preservation

Group 3 was just like Group 1, but Group 3 saw only 12 attributes. Our initial intent was that Group 3 would have the top 12 attributes, but our initial estimate (based on a sample of 10) was wrong. Group 4 saw a different set of 12 attributes. Groups 3 and 4 had the minimal overlap of 4 attributes.

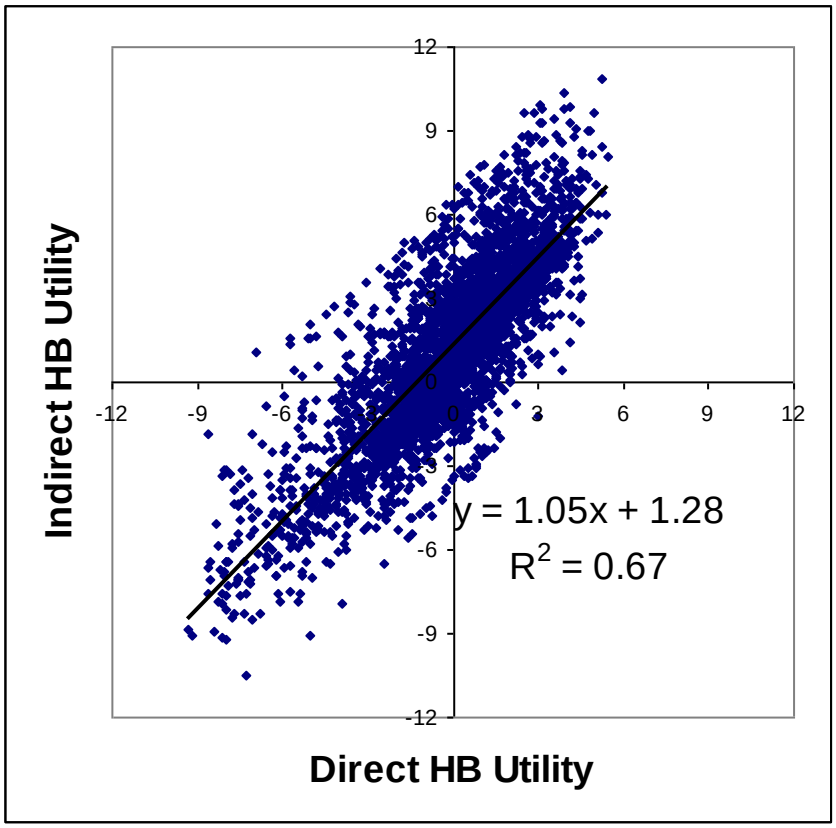
The objective in showing subsets was to test what happened when the anchoring went from all 20 attributes to a subset of 12. In theory, the anchoring should be the same whether respondents saw 12 attributes or 20. In practice, respondents are known to contextualize their responses, and indeed this is what we observed here.

First we noticed that respondents doing the direct approach (which showed 10 attributes on a screen twice), were more critical. That is, these respondents were less likely to say an attribute was Very Important. The table below shows that only 4-5% of respondents clicked an attribute as Very Important in the Direct method, but did not say it was Very Important in the Indirect method. In contrast, about 18-20% of respondents said an attribute was Very Important in the Indirect method but did check it as important in the Direct method. So the check marks definitely indicate a more critical attitude for the Direct approach, at least when 10 attributes are shown per screen.

Repercentaged

Direct Grid	Indirect Grid	4 Att MD/ 10 per Grid	5 Att MD/ 10 per Grid	4 Att MD/ 10 per Grid	5 Att MD/ 10 per Grid
Match Sign		64.60%	60.50%	75.80%	76.70%
Positive	Negative	3.60%	3.70%	4.20%	4.70%
Negative	Positive	17.00%	14.70%	19.90%	18.60%
Pos or Neg	No Info/ Inconsistent	14.80%	21.20%		

This more critical attitude in the Direct method toward which attributes are Very Important is confirmed with the scatterplot of the utilities. In the scatterplot below, each point is the utility for a specific respondent on a specific attribute, showing the utilities from both methods.

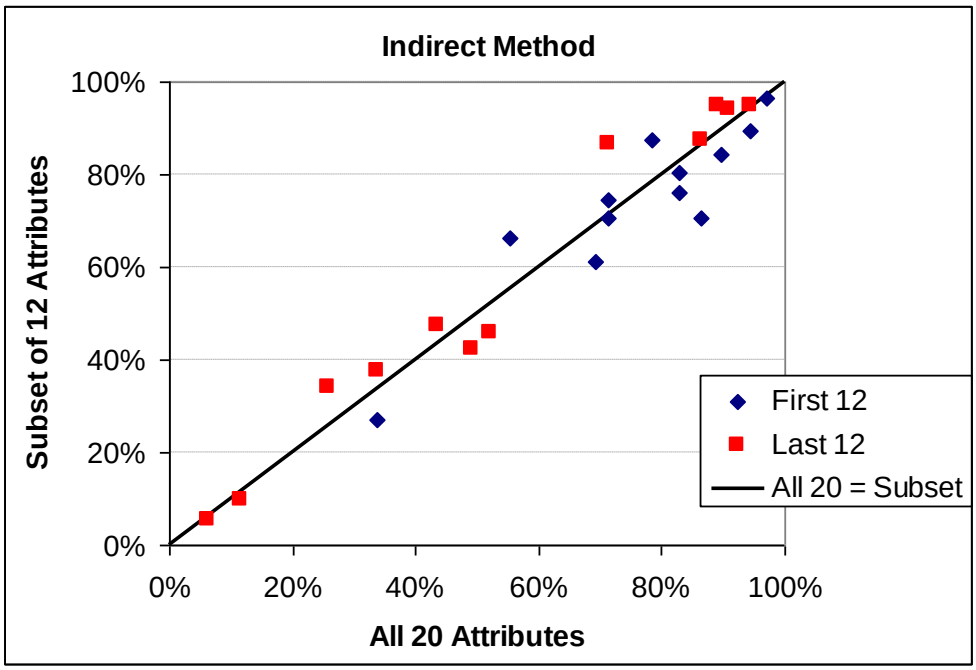


Utilities for the Indirect Dual Response Augment are shifted more positively. So we see that there is a difference between the two, but why?

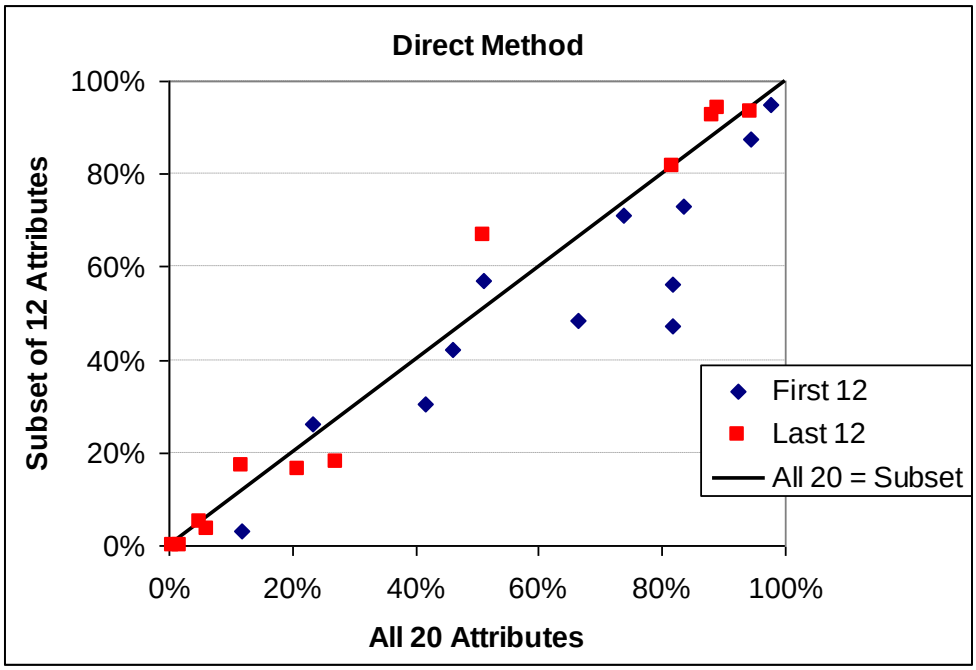
One potential explanation for this is that respondents who see 10 attributes on a screen are comparing all 10 attributes to each other and their Very Important grade is based on these that are Very Important compared to the others. In contrast, with the indirect augment respondents only saw four or five attributes on a screen, and were doing less comparative work to assess whether an attribute was Very Important.

This context sensitive explanation becomes even more plausible when we consider Groups 3 and 4, where only 12 of the 20 attributes were shown. If respondents did not apply contextual relativity then we would expect the two 12 attribute subgroups to be similar to the results from when all 20 attributes are shown.

The scatterplot below shows the percentage of positive utilities for an attribute using the Indirect Method. The x-axis shows the percentage positive for Group 1 doing all 20 attributes. The y-axis shows the percent positive for Groups 3 and 4, who did a subset of 12 attributes. Ideally we would expect all the attributes to fall on or near the line, indicating the same percentage of positive utilities for an attribute whether all 20 were shown or just a subset of 12.



In contrast when we look at the Direct method we see more divergence from the diagonal line.



So the indirect method shows better preservation of the Very Important threshold when only a subset of attributes are shown. This means if one wants to adopt the method that is most likely to capture the absolute threshold one should use the Indirect Dual Response augment. The direct method introduces some contextual relativity – and will change more as the attributes change.

VII. Conclusion

The Indirect Dual Response Method will be indeterminately anchored for some attributes. This indeterminacy is excessive when the incomplete coding is used, and we showed how this led to completely unacceptable results. But even with the revised complete coding of the Indirect method, some indeterminacy occurs. This indeterminacy increases with the number of attributes shown per MaxDiff task, and as the threshold is more evenly distributed (50% of attributes are positive and 50% negative). For these reasons we recommend showing four attributes at a time with the Indirect method, and certainly no more than five attributes at a time. If one must show six or more attributes per MaxDiff task then we recommend the Direct method.

While the Direct method is more accurate in theory, real respondents tend to apply a contextual relativity in evaluating whether an attribute meets a threshold like "Very Important". If one can live with some degree of contextual relativity, then the Direct method is preferable. But if it is important to avoid this contextual relativity for the anchoring then one must weigh the importance of less context dependence against the indeterminacy of the Indirect method