MACRO
CONSULTING, INC.
Marketing Research / Advanced Analytics

An Examination of the Components of the NOL Effect in Full-Profile Conjoint Models

In this paper, we confirm the existence of both algorithmic and psychological components of NOL for full-profile metric conjoint, examine the time decay of the psychological component and further develop a solution originally proposed in McCullough (1999) to completely eliminate NOL effects from full-profile trade-off models.

Author: Paul Richard “Dick” McCullough

2000 Sawtooth Software Conference Proceedings,
March 2000, Hilton Head Island, SC.

An Examination of the Components of the NOL Effect in Full-Profile Conjoint Models

Abstract

The existence of the number of levels effect (NOL) in conjoint models has been widely reported since 1981 (Currim et al.). Currim et al. demonstrated that the effect is, for rank-order data, at least partially mathematical or algorithmic. Green and Srinivasan (1990) have argued that another source of this bias may be behavioral. Although NOL can significantly distort study findings, no method for eliminating NOL, other than holding the number of attribute levels constant, has been discovered.

In this paper, we confirm the existence of both algorithmic and psychological components of NOL for full-profile metric conjoint, examine the time decay of the psychological component and further develop a solution originally proposed in McCullough (1999) to completely eliminate NOL effects from full-profile trade-off models.

INTRODUCTION

The existence of the number of levels effect in conjoint models has been widely reported since 1981 (Currim et al.). The effect occurs when one attribute has more or fewer levels than other attributes. For example, if price were included in a study and defined to have five levels, price would appear more important than if price were defined to have two levels. This effect is independent of attribute range, which also can dramatically affect attribute relative importance.

NOL was originally observed for rank-order preferences but has since been shown to occur with virtually all types of conjoint data (Wittink et al. 1989). Currim et al. demonstrated, for rank-order data, that the effect is at least partially mathematical or algorithmic. Green and Srinivasan (1990) have argued that a source of this bias may also be behavioral. That is, attributes with higher numbers of levels may be given more attention by respondents than attributes with fewer levels. If true, this might cause respondents to rate attributes with a greater number of levels higher than attributes with fewer levels. Steenkamp and Wittink (1994) have argued that the effect is, at least partially, due to non-metric quality responses, which computationally causes ratings data to behave similarly to rank-order data.

The NOL effect behaves somewhat differently for rank-order data and metric data. No NOL effect has so far been detected by simply removing levels from metric data in Monte Carlo simulations. However, there appears to be some question of whether or not there can be an algorithmic component of NOL for metric data derived from human responses, if the assumptions of normal, independent error terms are not met.

On the other hand, for rank-order data, it has been widely reported since Currim et al. that an NOL effect can be detected that is at least partially algorithmic by removing levels. Thus, in this strict sense of algorithmic component, the NOL effect from rank-order data may have both an algorithmic and psychological component but the NOL effect from metric data may have only a psychological component. The question is still open as to whether or not an algorithmic component for metric data exists when the data are derived from human responses.

It is generally agreed that the NOL effect is a serious problem that can and often does significantly distort attribute relative importance scores, utility estimates and market simulation results. And largely due to the fact that the only known method for removing this effect has been to hold the number of levels constant across attributes, it has often been ignored in commercial studies. McCullough (1999) suggested an approach that may eventually prove practical in eliminating NOL effects in full-profile conjoint studies. This paper further develops the concepts originally proposed there.

METHODOLOGICAL OBJECTIVES:

The objectives of this paper are:

- For full-profile metric conjoint, confirm (or deny) the existence of and estimate the separate magnitudes of the algorithmic and psychological components of NOL
- Confirm (or deny) the existence of and estimate the magnitude of the order effect potentially present in the two-stage conjoint approach (see McCullough (1999))
- Measure the effect of time on the psychological component of NOL
- Quantify the learning effect of exposure to level specifications prior to conjoint exercise
- Suggest a potential solution to eliminate NOL
- Validate the key assumption of that solution, i.e., that the psychological component diminishes rapidly over time when viewed in conjunction with an order effect

RESEARCH OBJECTIVE:

The objective of the study is:

- Identify key drivers in Web survey banner ad solicitations

STUDY DESIGN:

Overall, a multi-cell study design has been constructed to isolate the effects of several potential biases to trade-off models, using a web survey for data collection. The potential biases addressed by this study are:

- Algorithmic component of NOL
- Psychological component of NOL

- A time-lagged effect of the psychological component: exposure during a conjoint exercise to an attribute with a large number of levels may have a lingering psychological effect on subsequent conjoint exercises that contain that attribute
- An order effect: in a two-stage conjoint study, that is, a study with two separate conjoint exercises, the existence of one exercise prior to the second may create an order bias
- A learning effect: exposing respondents to attribute levels prior to a conjoint exercise may create a bias which is referred to here as a learning effect

To be able to analytically isolate and measure the magnitude of each of the above effects the study was split into four cells. The survey outline for each cell is as follows:

- Cell1= DQ || F, 2, demo's
- Cell2 = DQ || 2, F, demo's
- Cell3 = DQ, 2&F mixed, demo's
- Cell4 = DQ || F, demo's, 2

Where:

- DQ denotes direct questioning to identify exterior levels of incentive attribute,
- || denotes a two-day pause between stages (the assumption being that learning effect can be eliminated by delaying subsequent stages),
- 2 denotes 2-levels, that is, exterior levels trade-off, a trade-off exercise containing only the two exterior levels of each attribute,
- and F denotes full-levels trade-off, that is, a trade-off exercise containing all levels of all attributes.

The data collection protocol was as follows:

- Email invitation to split-cell web survey:
  Potential respondents were invited to participate in the survey via email.

- nth respondent receives survey to cell n mod 4:
  Every respondent that came to the survey website was routed through a counter which assigned respondents to each of the four cells in rotating order.

- Sample frame generated from email panel:
  Sample frame was purchased from an email list supplier. Opt-in names only were purchased. Opt-in lists are comprised of people who have previously agreed to allow themselves to be contacted for surveys. A small portion of the sample was obtained from the Bauer Nike Hockey website where visitors to that site were invited, via a banner ad, to participate in the survey.

- Metric full-profile conjoint study conducted via web survey:
  The trade-off exercises were designed using Sawtooth Software's CVA program. Conjoint measurement was pairwise ratings on a 9 point scale. There were 20 paired ratings in the full-levels trade-off exercises (D efficiency = 95%) and 4 paired ratings in the two-levels trade-off exercises (D efficiency = 100%). Individual level utilities were estimated for both full and exterior-only exercises using Sawtooth Software, Inc.'s CVA software.

- Sample size:
  Approximately 8,500 potential respondents were invited to participate. Roughly 70% of those completed the direct questioning segment of the survey and roughly 40% of those returned and completed the second segment of the survey. In cell 3, there was no time delay between the direct questioning segment and the remainder of the survey. Consequently, initial sample size for cell 3 was 1,474. Sample sizes in all cells were reduced to those whose direct questioning exterior levels matched perfectly their derived exterior levels (see Analysis section below). Roughly 22% of the completed surveys (174 per cell, on average) in all cells had matching direct questioning exterior levels and derived exterior levels. Additionally, samples were rescreened to eliminate only those respondents whose derived utility weights for their claimed exterior levels were statistically significantly different from their derived exterior levels. For these statistically matched samples, average sample size was approximately 600, or 85% of initial completes. Both data sets are discussed in the Analysis section below.

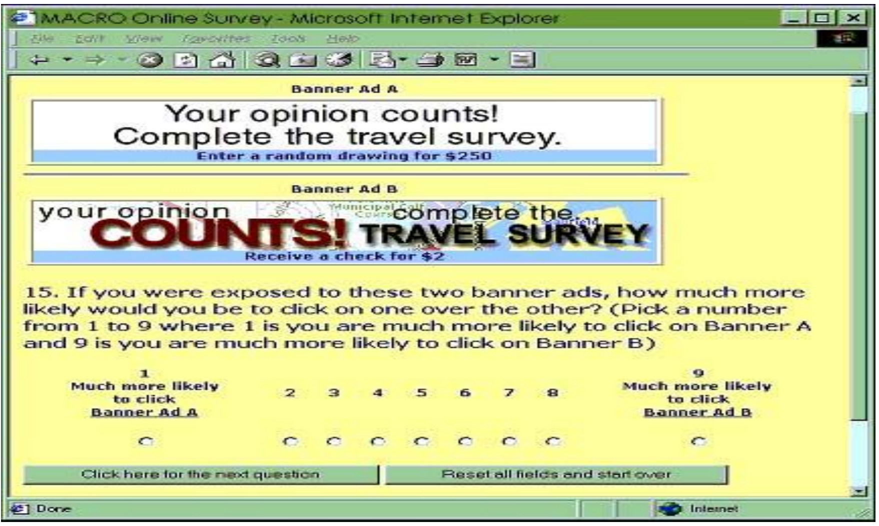
Data collection took place November 29 through December 18, 1999.

Trade-off attributes used in all eight conjoint exercises and the levels used in the four full-levels conjoint exercises were:

- Text only vs. graphics and text (2 levels)
- Animation vs. static (2 levels)
- Incentive (9 levels):
  - Random drawing for $250 cash
  - Random drawing for $2,500 cash
  - Random drawing for an Italian leather briefcase
  - Random drawing for a week in Hawaii for two
  - Each respondent receives a check for $2
  - Each respondent receives a check for $25
  - Each respondent receives a letter opener
  - Each respondent receives a Swiss Army knife
  - No incentive

All respondents were directed to a URL which contained all or part of the web survey for the cell they were assigned to. An example of the screen shown for a typical pairwise ratings question from one of the conjoint exercises can be found in Figure 1.

Figure 1


MACRO Online Survey - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Done
Internet

Banner Ad A

Your opinion counts!
Complete the travel survey.
Enter a random drawing for $250

Banner Ad B

your opinion COUNTS! complete the... TRAVEL SURVEY
Receive a check for $2

15. If you were exposed to these two banner ads, how much more likely would you be to click on one over the other? (Pick a number from 1 to 9 where 1 is you are much more likely to click on Banner A and 9 is you are much more likely to click on Banner B)

1 Much more likely to click Banner Ad A	2	3	4	5	6	7	8	9 Much more likely to click Banner Ad B
○	○	○	○	○	○	○	○	○

Click here for the next question | Reset all fields and start over

Note: 36 versions of the two-levels design were needed (nine levels taken two at a time). Each respondent saw the version which contained his/her exterior levels for the incentive attribute.

Let Cell1(2) = relative importances from 2-levels conjoint in cell1 and Cell1(F) = relative importances from full-levels conjoint in cell1, similarly for all other cells.

And let:
A = Algorithmic component,
P_I = Psychological component at time I
Oij = Order effect when conjoint exercise i precedes conjoint exercise j
L = Learning effect (due to exposure to levels during direct questioning)

Note that there are three P_I: P_0, P_1 and P_2. P_0 is the psychological component of NOL during a trade-off exercise that contains unequal numbers of levels across attributes. P_1 is the psychological component of NOL immediately after a trade-off exercise that contains unequal numbers of levels across attributes. P_2 is the psychological component of NOL a brief time after a trade-off exercise that contains unequal numbers of levels across attributes. Thus, in cell 1, where the full-levels trade-off is followed immediately by the two-levels trade-off, P_1 is the form of psychological component that would be affecting the two-levels trade-off. In cell 4, where the full-levels trade-off is followed by a demographics battery and then the two-levels trade-off, P_2 is the form of psychological component that would be affecting the two-levels trade-off.

Also, there are two forms of O_{ij} potentially at work: O_{F2} and O_{2F}. O_{F2} is the order effect when full-levels precedes two-levels. O_{2F} is the order effect when two-levels precedes full-levels.

Each of these eight different trade-off exercises (with the exceptions of Cell1(F) and Cell4(F)) will have a different combination of these various sources of bias operating. Table 1 below summarizes the sources of bias operating on each of the different trade-off exercises.

Table 1.

Cell	Bias Sources
Cell1(F)	A and P_0
Cell1(2)	O_{F2} and P_1
Cell2(F)	A, P_0 and O_{2F}
Cell2(2)	nothing
Cell3(F)	L, A and P_0
Cell3(2)	L and P_0
Cell4(F)	A and P_0 (same as Cell1(F))
Cell4(2)	O_{F2} and P_2 (similar to Cell1(2))

Jayme Plunkett and Joel Huber have both verbally expressed the opinion that the psychological effect may be short term . So short term that it may not appear or at least not appear fully when full-levels precedes two-levels in a two-stage design. If so, and if order effect is negligible, then it seems a good solution to NOL would be to do full-levels, calculate utils on the fly, insert derived exterior levels into a 2-levels conjoint (all with same respondent and within the same interview) and omit direct questioning altogether. This would avoid the discarded sample problem discussed in McCullough (1999). In cells 1 and 4, varying amounts of demo's (from none to some) have been inserted between full-levels and 2-levels to measure the effect of time on the psychological effect, if it is indeed short-term.

If Cell1(2) = Cell2(2), then $P_1 + O_{F2} = 0$, the psychological/order component is very short-term and the on-the-fly solution should be viable. Note that this argument assumes that P_1 and O_{F2} have the same sign.

If Cell1(2) !Cell2(2) but Cell4(2) = Cell2(2), then $P_2 + O_{F2} = 0$, the psychological/order component is short-term, but not very short-term, and the on-the-fly solution should be viable, if the second trade-off exercise is delayed for a short amount of time by the insertion of other survey questions, such as a demographics battery. Again note that this argument assumes that P_2 and O_{F2} have the same sign.

The above design and analysis should allow us to:

- Confirm (or deny) the viability of the on-the-fly solution
- Isolate and estimate magnitudes for A, P_0, O_{2F} and L
- Measure the time decay of P and O in combination, that is, measure $P_1 + O_{F2}$ and $P_2 + O_{F2}$

ANALYSIS

In each cell, respondents were asked directly what were their most and least preferred incentive levels. These claimed exterior levels were used in the two-levels trade-off exercise. Both of the other attributes had only two levels so no direct questioning was required to identify their exterior levels. Respondents whose claimed exterior levels, based on direct questioning were different from their derived exterior levels, based on the full-levels trade-off exercise, were excluded from this analysis. Recall from above that "different" is defined two ways: not perfectly, i.e., numerically exactly, matched and statistically significantly different.

Table 2 shows the frequency and incidence breakdowns by cell, for both perfectly matched and statistically matched samples.

Table 2.

Cell	Invitations	1st part completes	2nd part completes	Perfectly Matched	Statistically Matched
1	2,000	1,386/69%	536/39%	138/26%	447/83%
2	2,000	1,359/68%	544/40%	127/23%	462/85%
3	2,500	1,474/60%	1,474/100%	286/19%	1,224/83%
4	2,000	1,374/69%	546/40%	144/26%	476/87%

Attribute relative importance scores were calculated for each attribute within each trade-off exercise for each individual by taking the absolute difference between the attribute level with the highest utility weight and the attribute level with the lowest utility weight (for the same attribute), summing the absolute differences across all attributes in the trade-off exercise, dividing that sum into each absolute difference and multiplying by 100. These individual attribute relative importance scores were then averaged across all respondents.

To measure the magnitude of various sources of bias, mean attribute relative importance scores for the incentive attribute are differenced. Since the other two attributes have only two levels, any NOL-related effect will be reflected entirely in the attribute relative importance scores for the incentive attribute.

Using this difference, the various bias sources can be estimated. For example, the magnitude of the algorithmic component of NOL, i.e., A, is defined as the attribute relative importance score for the incentive attribute in Cell3(F) minus the attribute relative importance score for the incentive attribute in Cell3(2), since Cell3(F) is affected by L, A and P_0 and Cell3(2) is affected by L and P_0 (see Table 1). In Table 3 below, several bias sources are defined in terms of the cells of this study.

Table 3.

Source	Definition
A	Cell3(F) – Cell3(2)
P_0	(Cell1(F) – Cell2(2)) – (Cell3(F) - Cell3(2))
O_{2F}	Cell2(F) – Cell1(F)
L	Cell3(F) – Cell1(F)
$P_1 + O_{F2}$	Cell1(2) – Cell2(2)
$P_2 + O_{F2}$	Cell4(2) – Cell2(2)
°2[1]	Cell1(F) – Cell4(F)

Statistical significance of the differences in two sets of attribute relative importance scores has been tested using both anova and t-tests.

RESULTS

Table 4a lists the attribute relative importance scores for all attributes for the perfectly matched samples.

Table 4a: Perfectly Matched Samples.

	Cell1(n=138)		**Cell2(n=127)**		**Cell3(n=286)**		**Cell4(n=144)**	
	Full	*Exterior*	*Full*	*Exterior*	*Full*	*Exterior*	*Full*	*Exterior*
Text	7.17%	2.09%	6.12%	3.68%	6.06%	5.68%	6.59%	3.50%
Animation	8.47%	1.96%	5.19%	2.42%	6.64%	6.04%	7.22%	1.77%
Incentive	84.36%	95.94%	88.69%	93.90%	87.30%	88.28%	86.19%	94.73%

Table 4b lists the attribute relative importance scores for all attributes for the statistically matched samples.

[1] *is listed as a bias source for methodological confirmation purposes only. It is known that Cell1(F) and Cell4(F) have exactly the same biases operating on them regardless of what those biases are, since these two cells have been implemented exactly the same way. Therefore there should be no statistically significant difference in their attribute relative importance scores.*

Table 4b: Statistically Matched Samples.

	Cell1(n=447)		Cell2(n=462)		Cell3(n=1,224)		Cell4(n=476)	
	Full	Exterior	Full	Exterior	Full	Exterior	Full	Exterior
Text	8.95%	6.81%	6.75%	8.80%	11.00%	14.29%	8.82%	6.95%
Animation	9.06%	5.05%	7.21%	8.30%	7.82%	12.04%	8.46%	4.94%
Incentive	81.99%	88.14%	86.04%	82.89%	81.15%	73.67%	82.73%	88.10%

Based on these data, the following calculations of differences in incentive attribute relative importances were made:

Table 5a: Perfectly Matched Samples.

Source	Definition	Difference
A	Cell3(F) – Cell3(2) =	-.98 p.pts.
P_0	(Cell1(F) – Cell2(2)) – (Cell3(F) - Cell3(2)) -9.54p.pts.- (-.98) p.pts.=	-8.56 p.pts
O_{2F}	Cell2(F) – Cell1(F) =	4.33 p.pts.[3]
L	Cell3(F) – Cell1(F) =	2.94 p.pts.[3]
$P_1 + O_{F2}$	Cell1(2) – Cell2(2) =	2.04 p.pts.
$P_2 + O_{F2}$	Cell4(2) – Cell2(2) =	.83 p.pts.
°	Cell1(F) – Cell4(F) =	-1.83 p.pts.

Table 5b: Statistically Matched Samples.

Source	Definition	Difference
A	Cell3(F) – Cell3(2) =	7.48 p.pts.[3]
P_0	(Cell1(F) – Cell2(2)) – (Cell3(F) - Cell3(2)) -0.9 p.pts.- 7.48 p.pts=	-8.38 p.pts.[3]
O_{2F}	Cell2(F) – Cell1(F) =	4.05 p.pts.[3]
L	Cell3(F) – Cell1(F) =	-.84 p.pts.
$P_1 + O_{F2}$	Cell1(2) – Cell2(2) =	5.25 p.pts[3]
$P_2 + O_{F2}$	Cell4(2) – Cell2(2) =	5.21 p.pts.[3]
°	Cell1(F) – Cell4(F) =	0.74 p.pts.

ANALYSIS OF PERFECTLY MATCHED DATA

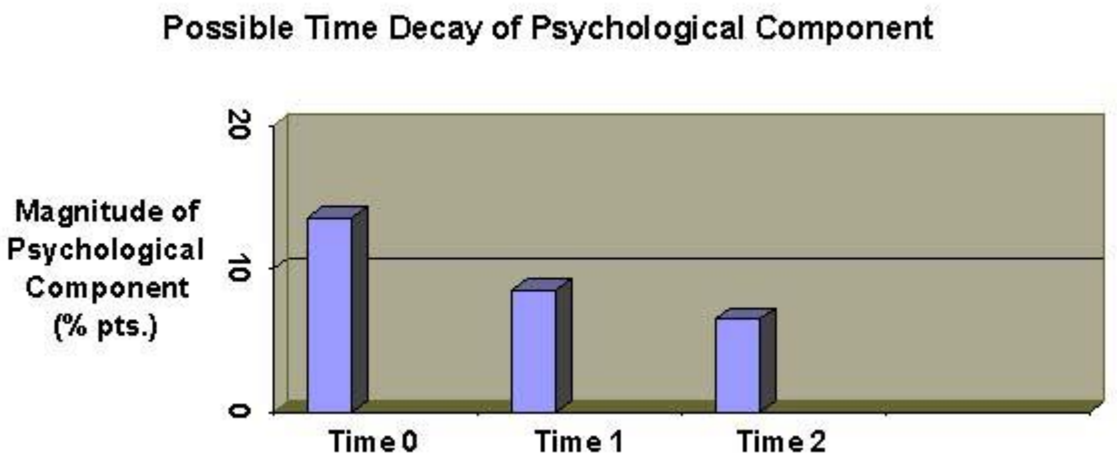
These data in table 5a show that, among this sample, there is a statistically significant psychological component of NOL but not an algorithmic component (at least not one detectable with these sample sizes). This psychological component is the largest of all the biases measured. Further, these data demonstrate that there is an order effect in the two-stage methodology that also significantly biases the attribute relative importance estimates. Also, there is a learning effect due to direct questioning that significantly biases the attribute relative importance estimates.

Intriguingly, these data also show that the combination of a time-delayed psychological effect and an order effect is negligible. This finding superficially suggests that a practical solution to eliminate the NOL effect is to do full-levels trade-off, calculate utils on the fly, insert derived exterior levels into a 2-levels trade-off (all with same respondent and within the same interview) and omit direct questioning altogether. Recall that this conclusion would require the assumption that P_1, P_2 and O_{F2} have the same sign. Given that fact that these data consistently and strongly suggest that the psychological component is negative and the order effect is positive, the validity of the two-stage approach as currently formulated must be questioned. That is, in this case, $P_1 + O_{F2} = P_2 + O_{F2} = 0$ but this finding does not generalize.

The fact that the attribute relative importances from Cell1(F) and Cell4(F) are statistically equal adds some face validity to the data collection process.

The data show that $P_1 + O_{F2}$ is statistically equal to zero. P_0 is known to be large and negative. If O_{F2} is roughly the magnitude of O_{2F} and P_1 is negative, then the magnitude of P_1 is roughly 8.5 percentage points. Similarly, P_2 would be roughly 6.5 percentage points. If true, this would allow us to chart the time decay of the psychological component. These data suggest such a chart might look like the one in Figure 2.

Figure 2.



As noted above, a careful review of Table 5a will show the surprising result that the psychological component is negative. That is, the large number of attribute levels in the incentive attribute cause incentive attribute relative importance to diminish, rather than increase. This result is consistent across all three cells which contain the psychological component P_0 in the full-levels exercise and not the two-levels, i.e., cells 1, 2 and 4. Recall that $P_1 + O_{F2}$ and $P_2 + O_{F2}$ were not statistically different from zero.

There are two possible explanations for this phenomenon. One possibility is that, in the full-levels exercise, respondents are exposed to the levels from both two-levels attributes, graphics and animation, roughly four and a half times more often than they are to the one nine-level attribute, incentive. This exposure may sensitize them to the two-level attributes, resulting in greater importance for the two-level attributes in the full-levels exercise and lesser importance for the incentive (nine-level) attribute. Conversely, in the two-level exercise, where attribute levels are all exposed equally, the two-level attributes would have less importance than in the full-levels exercise and the incentive attribute would have more.

Another possible explanation is that respondents, when faced in the two-levels exercise, with banner ads that have either their most preferred incentive or their least preferred incentive, give polarized responses. That is, respondents may have tended to give more 1 or 9 ratings because the incentive attribute either had a level respondents liked strongly or disliked strongly. This is all the more likely given the fact that the incentive attribute is overwhelmingly most important of all three attributes tested. This behavior may be an example of the utility balance issue discussed in Wittink et al. (1992a), Wittink et al. (1992b) and again in Wittink et al. (1997). That is, the incentive attribute may have greater attribute relative importance in the two-levels case because the utility imbalance is extreme in the two-levels trade-off. Wittink has demonstrated that utility imbalance will increase the magnitude of the NOL effect.

It is also possible that the sign of the psychological component may be a function of the data collection method. Perhaps, for example, monadic ratings inspire a different psychological component in respondents than pairwise ratings.

ANALYSIS OF STATISTICALLY MATCHED DATA

The statistically matched data offer somewhat different results from the perfectly matched data. Similar to the perfectly matched data, the statistically matched data show a statistically significant and negative psychological component and a statistically significant order effect (O_{2F}). However, the statistically matched data also show a statistically significant algorithmic component, statistically significant $P_1 + O_{F2}$ and $P_2 + O_{F2}$ and a statistically insignificant learning effect.

The fact that $P_1 + O_{F2}$ is statistically equal to $P_2 + O_{F2}$ implies that $P_1 = P_2$. And if we assume that O_{F2} is statistically equal to O_{2F} , we can conclude that $P_1 = P_2 = 0$. Thus, for the perfectly matched sample, the time decay of the psychological component of NOL appears to be slow while for the statistically matched sample, it appears to be quite rapid. This appears consistent with the fact that for the perfectly matched sample, there was a significant learning effect but for the statistically matched sample, there was not.

The fact that the attribute relative importances from Cell1(F) and Cell4(F) are statistically equal again adds some face validity to the data collection process.

DISCUSSION

In summary, we have two different data sets that yield somewhat different results. However, regardless of the way the data are analyzed, it must be concluded that there is a sizable psychological component that, surprisingly, can be negative and that there is a significant order effect (O_{2F}).

These data also suggest that a two-stage conjoint design where respondents do full-levels trade-off, utils are calculated on the fly, derived exterior levels are inserted into a 2-levels conjoint (all with same respondent and within the same interview) may be seriously flawed as a method of eliminating the NOL effect, due to the existence of a significant order effect (assuming $O_{2F} = O_{F2}$), unless some method of handling the order effect can be developed.

A final comment on matching respondents' claimed exterior levels (via direct questioning) to their derived levels. Across all cells, 22% of respondents had perfect matching. That is, 22% of respondents had claimed exterior levels that matched perfectly with the utility levels derived from the full-levels conjoint exercise. It appears possible that these respondents may be capable of more metric quality responses than those respondents who did not exactly match their claimed exterior levels and their derived exterior levels. If so, then the algorithmic component of the NOL effect measured with the perfectly matched data could be minimized by metric quality responses (refer to Steenkamp and Wittink). It may be the case that with a sample of less metric quality responses, such as the statistically matched samples, more statistically significant results would be obtained, i.e., an algorithmic component might be shown to exist, both because a larger NOL effect would exist and also due to the larger sample sizes. That is exactly what has occurred. Only the learning effect has diminished with the larger sample size.

Clearly, the negative psychological component is a surprising result. The fact that this result is consistently and clearly reflected in the data makes it hard to ignore. There are other results that are also puzzling:

- Why does an algorithmic component appear with the statistically matched data but not with the perfectly matched data?
- Why does the learning effect appear with the perfectly matched data but not with the statistically matched data?
- Why do $P_1 + O_{F2}$ and $P_2 + O_{F2}$ appear with the statistically matched data but not with the perfectly matched data?

One possible answer for the lack of algorithmic component among the perfectly matched sample may be that, for metric quality responses, the regression model error terms may not violate the assumptions of normality and independence. Conversely, it may be the case that the statistically matched sample generated non-metric quality responses and violated the error term assumptions.

The existence of a learning effect among the perfectly matched sample may again be influenced by non-metric quality responses. Would respondents capable of metric quality responses have greater recall of the direct questioning portion of the survey during the conjoint exercises and, therefore, be more influenced?

Interestingly, and perhaps related to the difference in learning effect results, perfectly matched samples appear to demonstrate a slower time decay of the psychological component than the statistically matched sample. Do metric quality respondents "remember" better? That is, does the psychological component of NOL decay more slowly for metric quality respondents than for non-metric quality respondents? Is the perfectly matched sample a sample of "super" respondents?

One possible explanation for the discrepancies between the results of the perfectly matched samples and the statistically matched samples is that the perfectly matched samples have been screened to retain only those respondents who are unusually "smart". They provide metric quality responses (and normal error terms), they remember the direct questioning experience and they retain the psychological influence longer (perhaps again because of better recall).

However, there are several other potential factors that may affect these results, as well:

- The incentive attribute was nominal, not ordinal or metric.
- Data collection was paired comparison rather than monadic ratings.
- The learning effect associated with the direct questioning stage of the survey may alter responses in some unanticipated way.
- The number of levels is radically different across attributes (two versus nine).
- The relative importance of the incentive attribute is overwhelmingly dominant.
- Fatigue: the full-levels exercise involved 20 cards while the two-levels exercise involved just four.

One argument against the appropriateness of using the statistically matched samples in analysis is that the statistically matched samples would have more noise, more error, making it more difficult to obtain statistically significant differences. But the statistically matched samples in this study found more statistically significant differences than the perfectly matched samples. Thus, this argument does not seem to have merit.

If metric quality responses are playing a role in these data, then it would appear that the statistically matched data sets would be more appropriate for analysis. If the learning effect associated with the direct questioning stage of the survey was also involved, then again it would appear that the statistically matched data sets would be more appropriate for analysis, since the statistically matched samples were not affected by a learning effect. The other factors: nominal attribute, paired ratings, number of levels disparity, relative importance disparity and fatigue, would apply equally to both the perfectly matched samples and the statistically matched samples.

Thus, it would appear that the statistically matched samples would be more appropriate for analysis and the conclusions derived from those data should be given greater weight than the conclusions derived from the perfectly matched data.

Based on these findings in combination with earlier studies, a clearer perspective on the NOL effect is beginning to emerge. The following hypothesis is consistent with existing literature:

> *There are two sources for the NOL effect: a psychological component due to disproportionate exposure to selected levels and an algorithmic component due to non-metric quality responses making the data act similar to rank order data. The psychological component is, at least sometimes, negative. In general, the algorithmic component is bigger than the psychological. In the study cited in this paper, the large number of levels of the most important attribute may have exaggerated the magnitude of the psychological component. In all other studies reported in the literature, the total NOL effect has been reported as positive. This result could be explained by an algorithmic component which is generally larger than the psychological component under more typical circumstances. None of these earlier studies had separated out the algorithmic component from the psychological component. Thus, a negative psychological component would have simply made the total observed NOL effect smaller but it would have remained positive.*

If the above hypothesis is true, then future studies should focus on how to remove each of the sources of the NOL effect separately. Perhaps the algorithmic component may be eliminated or minimized by asking questions in such a way that respondents are able to give metric responses. To combat the psychological component, perhaps there can be developed experimental design strategies that constrain each level of each attribute to be shown an equal number of times, without sacrificing the model's ability to estimate unbiased parameters. Another avenue for investigation is utility balance. As has been discussed earlier, Wittink has shown that by balancing total utility in pairwise ratings the NOL effect is diminished. Does utility balance affect the algorithmic component, the psychological component or both? The implication of a better understanding of the sources of the NOL effect is that we have new areas to examine for potential solutions.

SUMMARY

Both algorithmic and psychological components of NOL were confirmed to exist and quantified. The psychological component was shown to be negative, at least in this case. The psychological component also appeared to decay rapidly over time, for the more general statistically matched samples, assuming the two order effects, O_{2F} and O_{F2}, to be equal in magnitude.

A solution to the NOL effect continues to be an elusive target. While the two-stage approach remains potentially useful, it cannot yet be viewed conclusively as a valid method for eliminating NOL. It appears that there is an order effect inherent in the two-stage approach that must be accounted for. However, given the lack of learning effect demonstrated here (for the statistically matched samples), the solution proposed in McCullough (1999) may be viable if: 1) respondents are screened to have statistically matched claimed and derived exterior levels rather than perfectly matched claimed and derived exterior levels and 2) the order of the full-levels trade-off and the two-levels trade-off is rotated to minimize the order effect. The amount of lost sample not only diminishes dramatically with the alternative screening method but the sample derived may be more representative of the target population. This revised approach would not suffer from a learning effect or a time-lagged psychological component of NOL and order effect would be minimized. Further work needs to be done to verify that the time-lagged psychological component of NOL is zero, that is, confirm the assumption $O_{2F} = O_{F2}$, and understand the magnitude of the resulting order effect when the two trade-offs are rotated.

Additional work needs to be done to understand how different types of respondents may have different NOL effects, depending on the quality of their responses and, perhaps, even on their memory capacities or other mental attributes.

References

Currim, I.S., C.B. Weinberg, D.R. Wittink (1981), "The Design of Subscription Programs for a Performing Arts Series," Journal of Consumer Research, 8 (June), 67-75.

Green, P.E., and V. Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," Journal of Marketing, 54 (October), 3-19.

McCullough, Dick (1999), "The Number of Levels Effect: A Proposed Solution," 1999 Sawtooth Software Conference Proceedings, 109-116.

Steenkamp, J.E.M., and D.R. Wittink (1994), "The Metric Quality of Full-Profile Judgments and the Number-of-Levels Effect in Conjoint Analysis," International Journal of Research in Marketing, Vol. 11, Num. 3 (June), 275-286.

Schifferstein, Hendrik N.J., Peeter W.J. Verlegh, and Dick R. Wittink (1999), "Range and Number-of-Levels Effects in Derived and Stated Attribute Importances," an unpublished working paper, Yale School of Management.

Wittink, D. R., (1990), "Attribute Level Effects in Conjoint Results: The Problem and Possible Solutions," 1990 Advanced Research Techniques Forum Proceedings, American Marketing Association.

Wittink, D. R., J. C. Huber, J. A. Fiedler, and R. L. Miller (1992a), "The Magnitude of and an Explanation for the Number of Levels Effect in Conjoint Analysis," working paper, Cornell University (December).

Wittink, D. R., J. C. Huber, P. Zandan, R. M. Johnson (1992b), "The Number of Levels Effect in Conjoint: Where Does It Come From and Can It Be Eliminated?," 1992 Sawtooth Software Conference Proceedings, 355-364.

Wittink, D.R., L. Krishnamurthi, and D.J. Reibstein (1989), "The Effects of Differences in the Number of Attribute Levels on Conjoint Results," Marketing Letters, 1, 113-23.

Wittink, D. R., William G. McLaughlan, and P. B. Seetharaman, "Solving The Number-of-Attribute-Levels Problem In Conjoint Analysis", 1997 Sawtooth Software Conference Proceedings, 227-240.

The author wishes to thank Rich Johnson and Dick Wittink for their invaluable assistance with the design of this study as well as Jayme Plunkett and Jamin Brazil for their help in collecting and analyzing the data.

2000 SAWTOOTH SOFTWARE Conference Proceedings, March 2000, Hilton Head Island, SC.

MACRO
CONSULTING, INC.

We are an independent ***marketing research consulting firm*** *dedicated to helping you make the most informed, insightful marketing decisions possible. We specialize in technology, consumer, and new product research, and are well recognized for our* ***State-of-the-Art Research*** *techniques****.***

Ultimately, we provide more than just technical expertise. We focus on developing ***pragmatic solutions*** *that will have a positive impact on the profitability of our clients.*

CONTACT US:
Telephone: 650-823-3042
General Inquiries:
info@macroinc.com

Advanced Analysis Inquiries:
analysis@macroinc.com

richard@macroinc.com

www.macroinc.com