Sawtooth Software

RESEARCH PAPER SERIES

How Many Questions Should You Ask in Choice-Based Conjoint Studies?

Richard M. Johnson
and
Bryan K. Orme,
Sawtooth Software, Inc.
1996

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

How Many Questions Should You Ask In Choice-Based Conjoint Studies?

Richard M. Johnson and Bryan K. Orme
Sawtooth Software, Inc. (360/681-2300)

ART Forum, Beaver Creek, 1996

When planning a choice-based conjoint study, one must decide how many choice tasks to give each respondent. Too many may produce biased or noisy results, and too few will reduce precision. We re-analyze data from 21 commercial studies, conducted in several countries and languages, with widely varying product categories, to see how results depend on the number of tasks respondents are given.

Introduction

In choice-based conjoint analysis, one of the most frequently asked questions is "How many questions should I ask each respondent?" This is an important issue because:

- We know that if the interview is too long, respondents can get fatigued or bored, and their answers may be of little value.

- But, we are motivated to collect as much data from each respondent as possible to maximize the impact of each dollar spent on fieldwork.

We thought it would be interesting to re-analyze data from many commercial choice-based conjoint studies, from different product categories, with different kinds of respondents, and from different countries, to provide empirical answers to several questions:

- How many choice tasks should you ask each respondent? Is there more error in answers to later tasks?

- How much information is contributed by multiple answers from each respondent? Is it better to ask 1,000 respondents one question each, or to ask 100 respondents 10 questions each? How much better?

- Is there a systematic change in respondents' answers as the interview progresses? Do brand or price become more important? Do respondents become more or less likely to choose the "None" option?

- How long does it take respondents to answer choice questions? How long is an interview with a certain number of tasks likely to take?

- Should you ask for just the first choice for each set of concepts, or is it useful to ask for second choices as well?

Our Approach

Many researchers like choice data, because answering choice questions is relatively similar to what people do when purchasing products in the real world. However, choice studies are less efficient than other types of conjoint studies. Before each answer, the respondent should process information about how several concepts differ on several attributes. But the answer indicates only which concept is chosen, with no indication of strength of preference, as is available in ratings-based conjoint. And, of course, we seldom learn which other alternatives would have been acceptable, or the reasons for preference or rejection of any alternatives.

Unlike other conjoint methods, choice-based conjoint studies seldom provide enough information to estimate utilities for individual respondents. Instead, we usually combine data from many respondents to estimate average utilities.

Choice-based conjoint studies are of two types:

> Fixed designs. Studies with fixed designs are usually done by paper and pencil, often with a single version of the questionnaire, but sometimes with several versions. Within a version, all respondents get the same questions. By careful application of experimental design principles, and sometimes by combining data from several different versions, one can collect enough information to estimate utilities for a group of respondents.
>
> Randomized designs. Studies with randomized (or customized) designs are usually administered in computer-assisted interviews, where every respondent has a unique interview. Although the word "randomized" suggests "haphazard" or "uncontrolled," randomized designs can be of high quality and permit efficient estimation of aggregate utilities.

One of the benefits of randomized designs is that data can be aggregated question-by-question. For example, since each respondent sees a unique first question, one can use data from every respondent's first question to estimate a set of utilities based on only that question. Similarly, one can estimate utilities for everyone's second question, third question, or last question. Thus, randomized designs permit an examination of how utilities may change as respondents progress through the interview, as well as how the relative error level varies.

With aggregate analysis, we often blur the distinction between *sampling error*, which is reduced by including more respondents, and *measurement error*, which is reduced by having more data from each respondent. We know that sampling error decreases inversely with the square root of the sample size, but not much is known about how error

varies with the number of choice tasks. Is it better to have 1,000 respondents, each with one task, or to have 100, each with 10 tasks? If so, how much better? Questions like this can best be answered with data from randomized designs.

Our data sets used randomized designs produced by Sawtooth Software's CBC System. CBC designs are of high quality, satisfying the Huber-Zwerina criteria of orthogonality, level balance, and minimal overlap. CBC makes no attempt to produce utility balance, preferring not to make a tradeoff with orthogonality. Although our data are from randomized designs, *our conclusions apply equally to fixed designs*.

A total of 21 data sets were contributed by CBC users, including: Dimension Research, Inc., Griggs Anderson Research, IntelliQuest, Inc., McLauchlan & Associates, Mulhern Consulting, POPULUS, Inc., SKIM Analytical, as well as several end-users of choice data. The studies included a wide variety of product categories ranging from beverages to computers and airplanes. They involved fieldwork done in several countries and languages. The number of attributes ranged from three to six, and the number of choice tasks ranged from 8 to 20. The numbers of respondents ranged from 50 to 1205, and altogether they contained approximately 100,000 choice tasks.

Because these data sets were not designed for methodological purposes, most did not include holdout tasks that could be used to assess predictive validity. Consequently, our analysis has centered around the topics of reliability and internal consistency.

Reliability

In this section, we investigate the similarity between logit solutions computed with *different samples of respondents*. Respondents in each data set were divided randomly into halves, and logit estimation was done separately for each half. We used effects coding, so the utilities for levels of each attribute summed to zero. The utility for "None" was expressed as one additional parameter. The reliability of each solution was measured by the square of the correlation coefficient computed across its attribute levels, as in the example in Table 1. The value of r^2 is often interpreted as the percentage of the variance in either set of estimates which is accounted for by the other, so 100 times one minus r^2 can be interpreted as a relative percentage of error.

Table 1
An Example of Correlation Between Sets of Utilities

	First Random Half of Sample	Other Random Half of Sample
Brand A	0.5	0.55
Brand B	-0.3	-0.31
Brand C	-0.2	-0.24
Package 1	-0.33	-0.3
Package 2	0.22	0.1
Package 3	0.11	0.2
Price 1	-1	-0.99
Price 2	0.1	0.01
Price 3	0.9	0.98
None	-0.05	0.08

r = .989; r2 = .978; Relative error = 2.2%

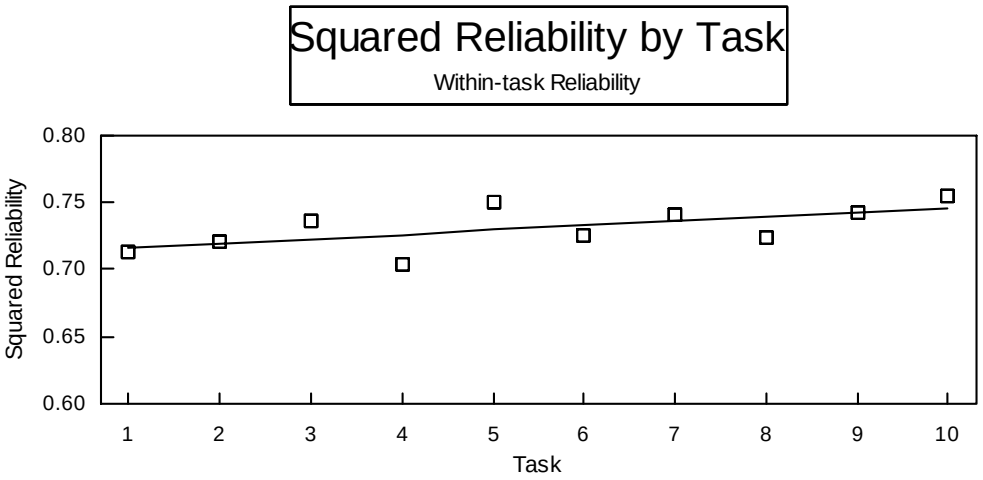
Within each data set, a squared reliability correlation was computed for each task separately, as well as for the first half of the tasks, for the last half of the tasks, and for all tasks. Finally, to make the results insensitive to a particular division of the sample, the entire process was replicated many times. We automated the process so it could run overnight and on weekends. The median number of replications was about 400 per data set. All in all, we computed about 400,000 logit solutions.

It is difficult to summarize task-by-task results for all of our data sets because they differ in numbers of tasks and numbers of respondents. In Table 2, we give average squared reliabilities for the first 10 tasks, averaged over the 8 data sets that had 300 or more respondents and 10 or more tasks. The squared reliabilities are also shown graphically in Figure 1.

Table 2
Average Squared Reliabilities for 8 Data Sets

Task #	Squared Reliability	Percent Error
1	0.713	
2	0.720	
3	0.736	
4	0.703	
5	0.750	
Avg. 1st Half	0.724	27.60%
6	0.726	
7	0.741	
8	0.724	
9	0.742	
10	0.754	
Avg. 2nd Half	0.737	26.30%

Figure 1



Contrary to what one might expect, later answers are a little *more* reliable than earlier ones, though not dramatically so. Reliability trends slightly upwards as the interview progresses, with no apparent decline for up to 10 tasks.

We have not attempted to provide standard errors for these results, or for any others in this presentation. Their computation would be very complex, since all of our results are concerned with trends and involve correlated observations. There is also a fundamental question about what we should take as our sample size; although these results are based on thousands of respondents, and hundreds of thousands of logit estimates, there are, after all, only 21 data sets. We think the diversity of our data permits valid inferences about choice-based conjoint results in general, but we will not make formal statements about statistical significance.

In Table 3, we show results for 20 data sets, comparing reliabilities for estimates using the first half of the tasks with those using the last half of the tasks. Indices greater than unity indicate that estimates based on data in the second half are more reliable.

Table 3
Squared Reliability for Second Half Indexed to First Half

# Tasks	# Resps	Index
8	356	1.00
	336	0.99
10	539	1.01
	420	0.99
	400	1.00
	399	1.12
	270	0.99
	170	1.00
	106	1.28
	92	1.07
12	1205	1.00
	1184	0.99
	136	0.98
	110	0.99
	50	1.25
15	251	1.03
	120	1.01
20	300	1.02
	300	1.00
	75	1.04
Average		1.04
Weighted Average		1.01

The index is greater than unity for 14 out of 20 studies, and its average is 1.04, indicating that estimates based on the second half of the data are slightly more reliable than those based on the first half. Because the largest departures from unity are for data sets with small sample sizes, we have also weighted by sample size, producing an average index of 1.01.

All our results suggest that there is a modest *increase* in reliability as we progress through the interview, at least for the first 20 tasks. The gain from respondents learning how to answer choice tasks seems to outweigh loss from fatigue and boredom, even for studies with up to 20 tasks. We were surprised by this finding, and gratified to learn that including many choice tasks isn't necessarily a bad idea.

We next examine the gain in precision obtained by increasing the number of choice tasks per respondent. Table 4 shows average squared reliabilities for half-length vs. entire questionnaires.

Table 4
Gains in Precision from Doubling the Number of Choice Tasks

# Tasks	# Resps	Avg. (Reliability)^2 Half	Avg. (Reliability)^2 Entire	Avg. % Error* Half	Avg. % Error* Entire	Relative** Error
8	356	0.797	0.869	20.30%	13.10%	0.645
	336	0.952	0.971	4.8	2.9	0.59
10	539	0.857	0.880	14.3	12	0.839
	420	0.974	0.979	2.6	2.1	0.791
	400	0.962	0.978	3.8	2.2	0.575
	399	0.858	0.871	14.2	12.9	0.911
	270	0.931	0.955	6.9	4.5	0.644
	170	0.935	0.950	6.6	5	0.769
	106	0.497	0.564	50.3	43.6	0.867
	92	0.449	0.500	55.1	50	0.907
12	1205	0.957	0.972	4.3	2.8	0.64
	1184	0.962	0.977	3.8	2.3	0.616
	136	0.882	0.919	11.8	8.1	0.689
	110	0.891	0.926	10.9	7.4	0.676
	50	0.492	0.636	50.8	36.4	0.716
15	251	0.860	0.905	14	9.5	0.679
	120	0.507	0.570	49.3	43.1	0.874
20	300	0.937	0.960	6.3	4	0.632
	300	0.932	0.955	6.9	4.5	0.657
	75	0.794	0.860	20.6	14	0.678
Average			0.720			
Weighted Average			0.690			

* Percent error = 100*(1-r2)
** Relative error = (%error for all tasks)/(%error for 1st half)

The final column indexes the percentages of error for estimation based on all tasks with respect to error for estimation based on half of the tasks. The average index is .72, indicating an average gain in precision of 28% from doubling the number of tasks. Because studies differ in sample size, we also provide the weighted average of .69, indicating a gain in precision of about 31% from doubling the number of tasks.

Because precision varies with the square root of sample size, doubling the number of *respondents* should produce a relative error of $1/\sqrt{2}$, or .707. That lies between our two estimates of the effect of doubling the number of *tasks*. Apparently, one gets approximately the same increase in precision from increasing the number of tasks as from proportionately increasing the number of respondents.

We believe this result should be interpreted carefully. Consider the absurd example of a single respondent with an extremely long interview. One would never use estimates from a single respondent to predict results for a heterogeneous population; we believe one must ensure a large enough sample size in choice-based conjoint interviews to adequately represent a population.

A further word needs to be said about long interviews. Although adding tasks appears to be an effective way to increase precision, and later tasks are apparently even more reliable than earlier tasks, we have not yet shown that later tasks measure the *same thing* as early tasks.

Do All Tasks Measure the Same Thing?

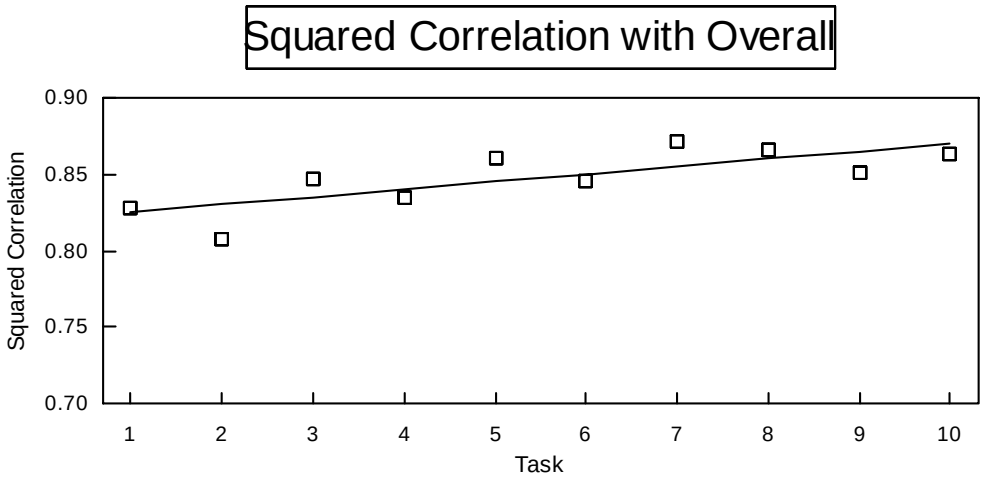
Rather than similarity of estimates from different *samples of respondents*, we now consider the similarity of estimates from different *tasks*. For each data set, we accept the estimates based on all tasks as best summarizing the information in the data set, and inquire how closely those results are approximated by estimates based on the first task alone, the second task alone, etc.

We first examine just those 8 data sets with 300 or more respondents and 10 or more tasks. In Table 5 we show average squared correlations between estimates based on each of the first ten tasks and estimates based on the entire data set, which are also displayed graphically in Figure 2.

Table 5
Average Squared Correlations with Overall Estimates

Task #	r^2
1	0.828
2	0.808
3	0.847
4	0.835
5	0.861
Avg. 1st Half	0.836
6	0.846
7	0.872
8	0.867
9	0.852
10	0.864
Avg. 2nd Half	0.860

Figure 2



Although the differences are again modest, later tasks are better predictors of results from the total interview, and estimates from the first two tasks are least like the total. This may support the practice of including warm-up tasks that are excluded from the analysis,

although all tasks have quite high r^2 values, suggesting that warm-up tasks may not really be necessary.

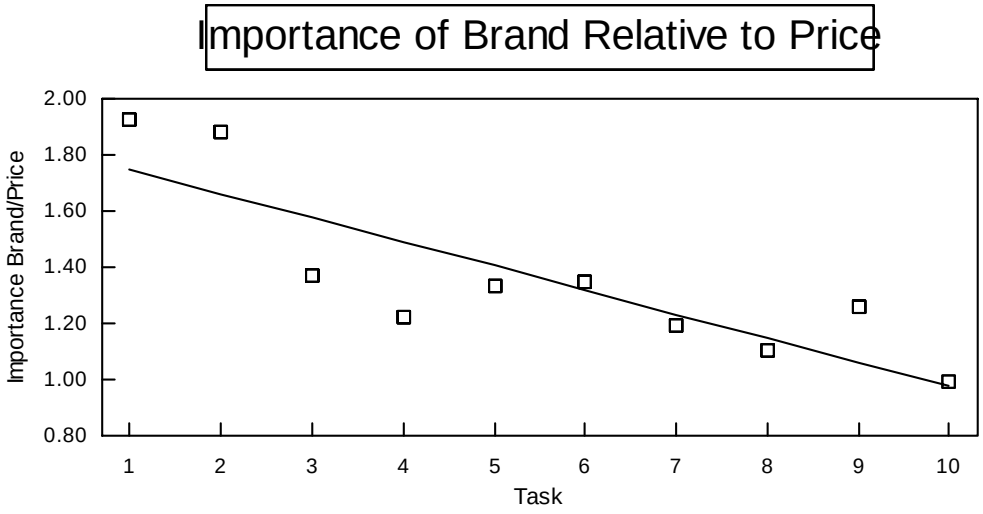
This upward trend in correlations suggests that some process may be occurring during the interview that makes results from the earliest tasks different from others. Previous experience had led us to believe that brand might diminish in importance as the interview progressed, and that price might become more important. Although the specific attributes and levels in our data sets were not identified, we had asked the contributors of our data sets to indicate which attributes corresponded to brand and price, if any. Both attributes were represented in 18 of our data sets. We examined task-by-task differences in importance of brand and price for 7 data sets that included 10 or more tasks and 300 or more respondents, as well as both brand and price.

Attribute importances were calculated in the usual way, by percentaging attributes' ranges in utility values. The importance of brand was divided by the importance of price for each task in each data set, and those indices were averaged across data sets. These results are shown in Table 6 and Figure 3.

Table 6
Average Importance of Brand Relative to Price

Task	Index of Importance (Brand/Price)
1	1.93
2	1.88
3	1.37
4	1.22
5	1.33
Avg. 1st Half	1.55
6	1.35
7	1.19
8	1.10
9	1.26
10	0.99
Avg. 2nd Half	1.18

Figure 3



As expected, there is a strong decreasing trend in these indices. Tasks one and two exhibit particularly strong brand importance relative to price. For choice studies with very few tasks, this effect could have a significant impact upon the overall measured importance of brand relative to price.

If we change our focus from task-by-task to an aggregate of first half vs. last half, we can expand this analysis to include the other data sets as well. Table 7a provides confirming results for 18 data sets.

Table 7a
Changing Importance of Brand vs. Price

# Tasks	# Resps	b2/b1	p2/p1	ratio
8	356*	0.72	1.22	0.59
	336*	0.83	1.08	0.77
10	539	0.72	1.19	0.6
	420	0.81	1.23	0.65
	399*	0.91	1.05	0.87
	270	0.96	1.07	0.9
	170	0.92	1.32	0.7
	106*	0.93	1.1	0.84
	92	0.87	1.11	0.78
12	1205	0.87	1.16	0.76
	1184	0.9	1.2	0.75
	136*	1.3	1.04	1.25
	110*	1.12	1.12	1
	50	0.38	1.12	0.34
15	251	1.09	1.11	0.98
20	300*	0.9	0.85	1.06
	300*	1.06	0.95	1.11
	75	0.85	0.97	0.88
		Average		0.82
		Weighted Average		0.79

*Choice interview preceded by another Conjoint interview

In Table 7a, we show the ratio b2/b1 of brand importance in the second half compared to the first half, and the ratio p2/p1 of price importance in the second half compared to the first half. Brand importance decreases in the second half of the tasks in 14 cases of 18, and price importance increases in 15 cases of 18.

In the final column, we show the ratio of the two previous columns, which shows the change in brand importance with respect to the change in price importance. All but four

of those values are less than unity, and their average is .82. Because some data sets have very small sample sizes, we have also provided the weighted average, .79. There is a net loss of about 20% in the importance of brand relative to price from the first half of the tasks to the second.

This table confirms the finding that the importance of brand decreases throughout the interview, and the importance of price increases. We regard this change in relative importance of brand and price as an important finding, and also as a reasonable one. Brand is often regarded as a proxy for other attributes, and it is not surprising that respondents apparently pay particular attention to brand in initial tasks. As they gain experience, they apparently become "better shoppers," learning to draw finer distinctions and becoming more sensitive to other attributes, including price.

We asked the providers of the CBC data sets to indicate whether another conjoint task (such as ACA) preceded the choice interview. Those studies are marked by asterisks next to the number of respondents in Table 7a above. Table 7b reports the shift in importance between brand and price between choice interviews that had preceding conjoint exercises versus those that didn't.

Table 7b
Changing Importance of Brand vs. Price
Effect of Preceding Choice with Another Conjoint Exercise

Ten data sets without preceding conjoint interview:

	b2/b1	p2/p1	ratio
Average	0.84	1.15	0.73
Weighted Average	0.87	1.17	0.74

Eight data sets with preceding conjoint interview:

	b2/b1	p2/p1	ratio
Average	0.97	1.05	0.94
Weighted Average	0.92	1.05	0.90

The summary data in Table 7b suggest that a warm-up conjoint interview considerably moderates the shift in importance between brand and price. When preceded by another conjoint interview, the shift in importance between brand and price is in the range of 5 to 10%. However, without a previous conjoint exercise, the shift is much greater--about 25% from the first to the second half of the interview.

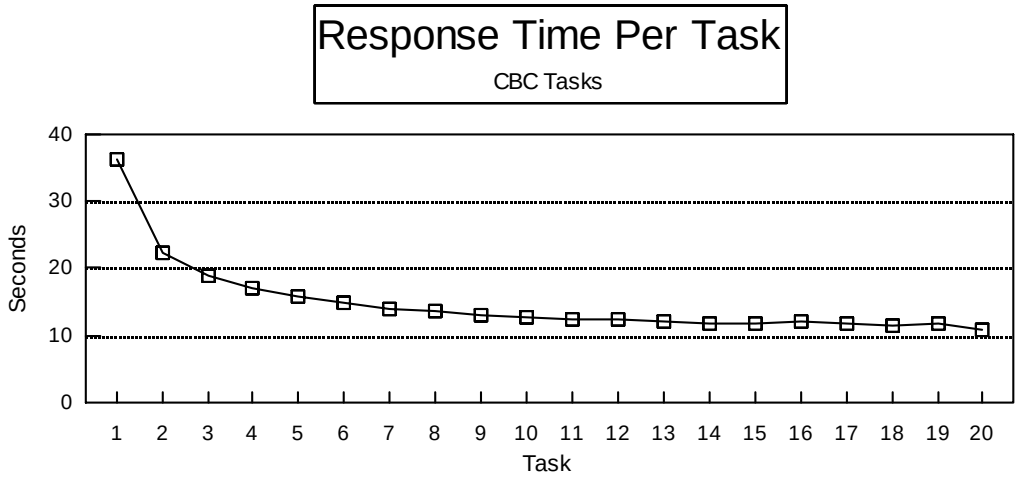
When there are different results from early and late parts of the interview, one is led to ask "which is correct?" We don't have the answer. One might argue that the very first task should be the best, since the respondent is less contaminated by the effect of previous questions. This seems likely to be true for impulse purchases. But, for real-world purchases in high-involvement categories, buyers probably spend more time considering the options, features, and pros and cons of each alternative. Later tasks seem to better reflect such behavior.

Interviewing Time

When deciding the number of tasks to ask, the time it takes to answer choice tasks can become an important consideration. CBC records the time in seconds for each choice, so it is easy to see how choice interviewing time depends on the number of tasks. We trimmed all times greater than 100 seconds to 100 before calculating the means described below. This prevents a respondent who takes a coffee break during a choice task from having a too-dramatic effect on average response times.

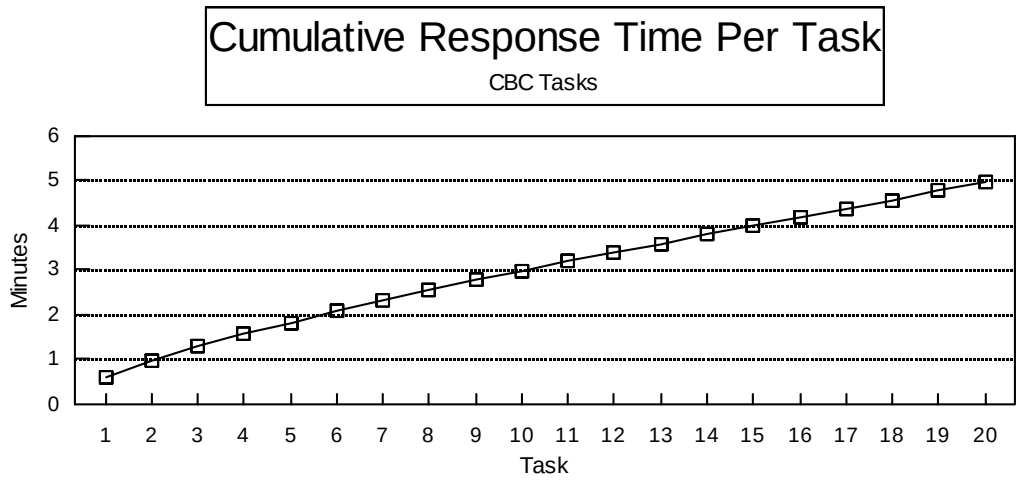
Figure 4 displays average response time per task and reveals that it takes respondents a couple of tasks to get "into the groove." This slow start is probably related to the higher error associated with the first few tasks.

Figure 4



Over the range of tasks we studied, response times gradually decrease as tasks increase until the last few choices take an average of only about 12 seconds to complete--about a third as long as the first task. Respondents are most productive in the last stages of many-task choice interviews, providing high quality data at a much faster rate. This finding supports the practice of having respondents complete many tasks. Figure 5 presents the same data as Figure 4, but in terms of cumulative time.

Figure 5



Choice interviews don't take much time; the average time even for as many as 20 tasks is only about five minutes. We observed average interview times ranging from a low of 1.5 minutes (a 12-task study with two concepts and a "None") to a high of 6.6 minutes (a 20-task study). Since we rarely ask more than 20 questions, choice interviews are usually less demanding than many traditional full-profile conjoint or ACA studies, which can last anywhere from 10 to 25 minutes.

We encouraged CBC users to provide us data sets with as many tasks as possible, so the 21 studies represented in this research are some of the longer CBC interviews being conducted today. Many CBC studies include just a few choice questions per respondent.

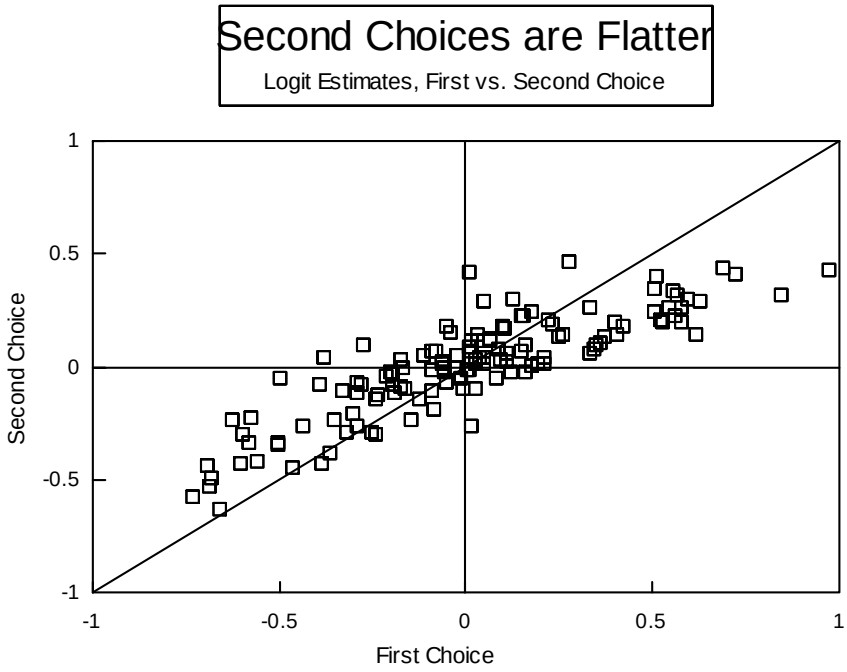
Second Choices

In addition to first choices, one may also collect second and subsequent choices among the concepts in each choice task. Six of our data sets included second choices.

Once a respondent has identified the most preferred concept, it requires little effort to record a second choice among remaining alternatives. In our data sets, second choices required only 14% as much time as first choices. This estimate may be a little low, since time spent identifying the second choice may spill over to first-choice time as, over repeated tasks, respondents learn they will be asked for both first and second choices. Even so, it is clear that much more data can be captured with little effort by including second choices.

If more data can be collected at such little cost, should budget-minded researchers use both first and second choices? To answer this, we compared logit coefficients estimated using first choices alone with those estimated from second choices alone to see if first and second choices differ. The scatter plot below shows considerable correlation between utilities calculated from first and second choices, with an r^2 of .76. We've added a line representing a slope of 1.0 to help illustrate that second choice estimates are flatter than first choice.

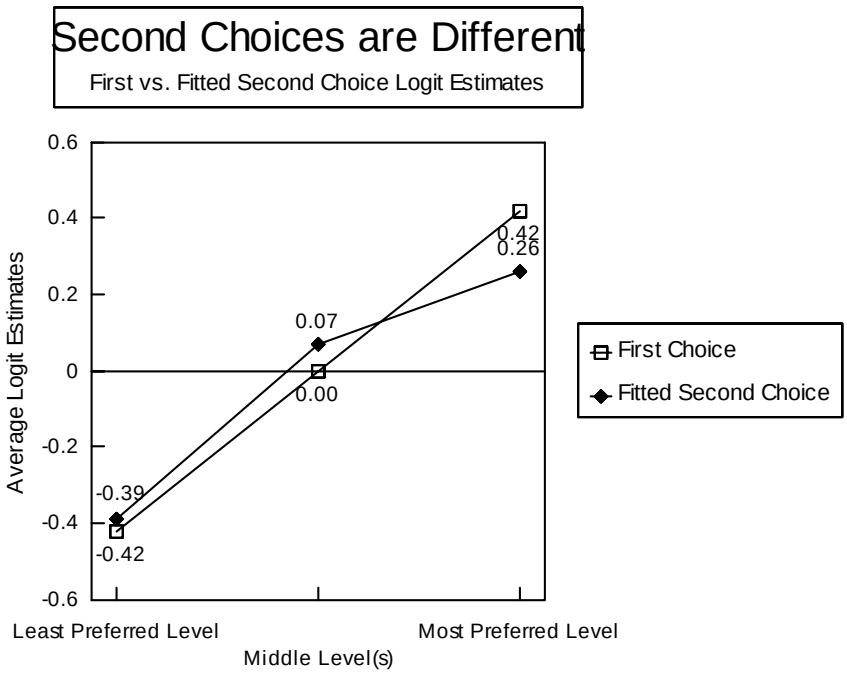
Figure 6


Second Choices are Flatter
Logit Estimates, First vs. Second Choice
Second Choice
First Choice
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5
1

The regression coefficient for predicting first choice utilities from second choice utilities is 1.36.

Since second-choice estimates are flatter than estimates from first choice, can we simply multiply them by a compensating factor before combining them with first choice results? We could if the flattening effect were the only difference, but all six studies with second choices display a disturbing pattern. The utility of the best level within each attribute is biased downward for second choice utilities. We have summarized this effect by averaging the logit estimates for the highest, middle, and lowest attribute levels (in terms of utility), over all attributes for first choices alone and second choices alone. We then fitted the second choice estimates to first choice by multiplying the second choice estimates by the slope which results in a least-squares fit, calculated on a study-by-study basis for each of the six data sets. The summary measures are displayed in Figure 7.

Figure 7



To determine the practical significance of this bias in second-choice utilities, we studied the residuals obtained when predicting first choice utilities from second choice utilities. Table 8 summarizes our findings for the most preferred level, middle level(s) and least preferred level for each attribute. The six studies we examined contributed a total of 34 attributes.

Table 8
Second-Choice Utilities Compared to First-Choice Utilities

	Average Residual (1st-2nd Choice)	Percent of Residuals that are Positive	Average t-value
Highest Level	0.162	30/34, or 88%	2.53
Middle Level(s)	-0.069	22/63, or 35%	-1.22
Lowest Level	-0.034	14/34, or 41%	-0.36

Where first and second choices lead to different conclusions, we believe that first choices better represent actual purchase conditions in the market, and for most product categories will produce better predictions of market behavior. The psychological context of second choices is different from first choices. It has been suggested that respondents may be trying to pick the best in a first choice, and concentrating more on avoiding the worst in subsequent choices. However, that does not seem to explain the fact that the preferred level of each attribute has lower estimated utilities for second choices.

We have suggested that the bias in second choices is a psychological, rather than an arithmetic artifact. To test this point, we used a Monte Carlo data generator which simulated a choice interview with three attributes, each with three levels, and each having pre-defined utilities. We simulated some degree of respondent error to the responses, and built a data set for 300 respondents with first and second choices and 16 tasks. After computing logit estimates for the data set, we saw no noticeable bias between first and second choice logit estimates for computer-generated data. This confirms our hypothesis that the bias in second choices results from some psychological process.

There is compelling evidence that utilities derived from second choices are different from those from first choices. We urge researchers to be aware that second choices can provide biased results.

Choice of "None"

It seems useful to permit respondents a "None" option, if only to add realism to choice-based conjoint tasks. However, sometimes "None" answers are used to infer how category volume would vary as products become more or less attractive. We are naturally interested in whether respondents are more or less likely to choose "None" as the interview progresses.

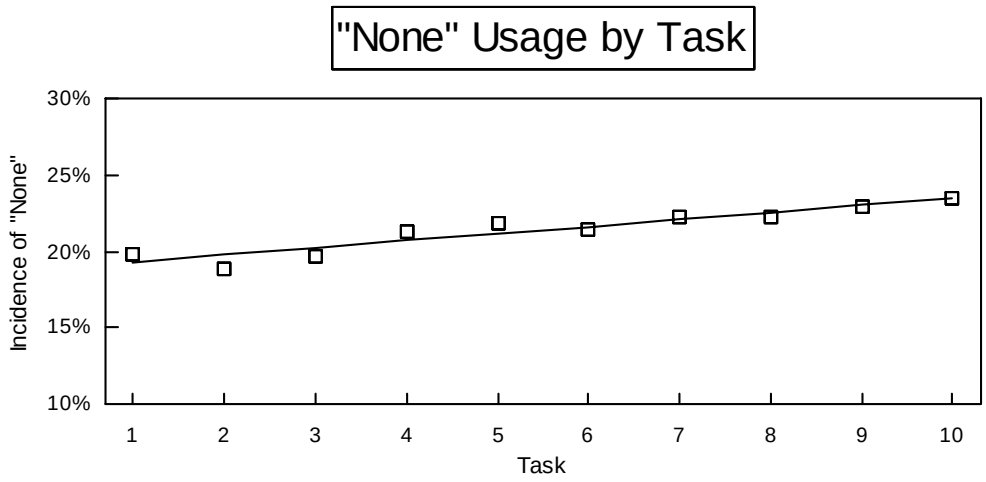
All of the data sets we received included a "None" option. On average, "None" was chosen 21% of the time over the first 10 tasks. We also found that respondents on average spent about ten percent less time for choices of "None" than choices of product concepts. Table 9 presents these findings by task, weighted by the number of respondents per study, and the incidence of Nones is presented graphically in Figure 8.

Table 9
None Usage By Task

	% Nones	None Time/ Other Choice Time
Task 1	19.8%	102.0%
Task 2	18.9	91.5
Task 3	19.7	90.5
Task 4	21.3	85.4
Task 5	21.9	84.4
Task 6	21.5	83.9
Task 7	22.3	88.5
Task 8	22.3	91.7
Task 9	22.9	92.4
Task 10	23.5	85.3
Average	21.4%	89.5%

Looking at the second column above, although results for the very first task are again a little anomalous, there doesn't appear to be much of a trend in the time spent making a "None" choice relative to other choices. (Though not shown in this table, we also examined tasks 11 through 20 and found no evidence that relative "None" times trended lower.) In contrast, "None" usage by task trends upward, and is illustrated in Figure 8.

Figure 8



Respondents are more likely to use "None" in later tasks. This could be due either to fatigue, or because respondents are reluctant to choose a mediocre concept after having seen better options in earlier tasks. Which is the case? We tend to believe the latter. If "None" were used principally by fatigued or bored respondents as a way to skip through the remaining choice questions, we would expect the relative "None" time to decrease in later tasks. Focusing only on studies which asked both first and second choice confirms

this finding. Fatigued respondents in a first and second choice study have an incentive to choose "None" as a first choice since this action skips the second-choice question. An examination of the percent "None" choices per task and relative time for Nones for the six studies with both first and second choice provided no indication that "None" was being used prevalently to avoid second-choice questions.

Reasons for Choosing None

There are at least two hypotheses for why respondents may choose "None."

> Economic Hypothesis: One hopes that respondents choose "None" to indicate that no offering is sufficiently attractive. If that is true, then "None" responses may be useful in forecasting how category volumes would change with differences in attractiveness of products.
>
> Decision Avoidance Hypothesis: Perhaps respondents choose "None" as a way of avoiding difficult choices. Previous analysis of other choice-based data had suggested that respondents may also choose "None" when the two most attractive concepts are nearly tied in utility. If that is true, then interview behavior would not reflect behavior likely to be found in real-world purchases, and "None" responses would be less useful in forecasting category volumes.

We examined this question for 15 data sets. For every respondent and choice task separately, we used utilities to estimate the probability that each concept would be chosen. The choice probabilities for the most attractive and second most attractive concept were noted, as well as whether "None" was chosen. We then did a regression analysis for each study with as many observations as there were respondents times choice tasks per respondent, where the independent variables were the attractiveness of the best and second-best concepts, expressed as choice probabilities, and the dependent variable was 1 if "None" was chosen and 0 otherwise.

We expected the regression coefficient for the most attractive concept to be negative, since we expect people to choose "None" more often when the best alternative is less favorable. However, we were more interested in the sign of the regression coefficient for the second-best concept. If it was also negative, that would suggest that choice of "None" is genuinely a response to an inferior set of options. However, if the second-best concept had a positive coefficient, that would indicate that choice of "None" increases as the second-best concept improves and the task becomes more difficult.

Among the 15 data sets examined, 8 had small coefficients for the second-best concept, with absolute values not exceeding twice their standard errors. For those data sets we cannot conclude anything about reasons for choice of "None." For the remaining 7 data sets, coefficients for the second-best concepts exceeded twice their standard errors, and all of them had negative signs. Thus, in those cases where there is any evidence about why respondents choose "None," all evidence favors the economic hypothesis over the decision avoidance hypothesis. We believe this finding lays to rest the conjecture that "None" responses are often used as a way to avoid difficult choices.

Our analysis argues that "None" usage appears to be a rational decision event. Even so, we caution against its use for volumetric demand estimates. Although this paper offers no evidence on this subject, our previous experience leads to us believe that respondents are poor at predicting their own purchase likelihood.

Summary

Although it will never be possible to produce guidelines that will be appropriate in all circumstances, we have learned enough from this analysis to provide suggestions that will be appropriate in most cases. We end by repeating the questions with which we started, as well as their answers:

> How many choice tasks should you ask each respondent? You can usually ask at least 20 choice tasks without degradation in data quality. Within that range, there is no evidence of increasing random error. Later tasks not only provide data at least as reliable as earlier tasks, but they are completed much faster by respondents.
>
> How much information is contributed by multiple answers from each respondent? Although there is no disputing the value of sample size, considerable gains can also be made from increasing the number of tasks per respondent. Within the ranges we studied, doubling the number of tasks per respondent is about as effective in increasing precision as doubling the number of respondents.
>
> Is there a systematic change in respondents' answers as the interview progresses? Do brand or price become more important? Do respondents become more or less likely to choose the "None" option? Yes to all three. Brand becomes less important, and price more so, and respondents are more likely to choose "None" as the interview progresses. These systematic effects are what influence the number of tasks each respondent should be given, rather than anticipated increases in random noise.
>
> How long does it take respondents to answer choice questions? How long is an interview with a certain number of tasks likely to take? Choice-based conjoint interviews go quite quickly. Average response times ranged from about 35 seconds for the first task to 12 seconds for the last. Even for 20 tasks, the longest average interview time was just under 7 minutes.
>
> Should you ask for just the first choice for each set of concepts, or is it useful to ask for second choices as well? Second choices provide more information at less cost, but they are biased. We advise asking only first choices.

We are surprised by some of these findings. There are three main things that we've learned from this analysis:

- Before doing this study, we were more concerned about burdening respondents with long choice questionnaires than we needed to be. We now realize that longer interviews can provide good quality information, though one must be aware that differences in interview length can produce shifts in brand/price tradeoffs.

- We had been impressed by the efficiency of asking for second choices, without adequate recognition of the bias inherent in their use.

- We had incorrectly suspected respondents often chose "None" to avoid difficult tasks, rather than because the offerings weren't attractive.

Fortunately, none of these surprises consists of bad news, and we think there is good reason for the enthusiasm with which choice-based conjoint analysis has been accepted by the market research community.