Sawtooth Software

RESEARCH PAPER SERIES

Fine-Tuning CBC and Adaptive CBC Questionnaires

Bryan Orme,
Sawtooth Software, Inc.

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

Fine-Tuning CBC and Adaptive CBC Questionnaires

Bryan K. Orme, Sawtooth Software

The Importance of Split-Sample Experiments

Among Sawtooth Software users, CBC (Choice-Based Conjoint) is the most commonly used conjoint-related technique (Sawtooth Software 2008). The newcomer, Adaptive CBC (ACBC), is also beginning to be used regularly. At Sawtooth Software, we look forward to research-on-research studies featuring split-sample tests that investigate ways to improve the effectiveness of questionnaires and how we analyze the data. Split-sample tests are often reported at the Sawtooth Software Conferences. We also are active in conducting such tests, and this paper reports on an experiment designed to test a variety of design and analysis issues for CBC and ACBC.

Western Wats with its Opinion Outpost Panel has been very helpful to us across multiple methodological studies like this, and we thank them for their excellent services. We conducted a web-based study among about 650 respondents related to preferences and features of fast-food restaurant drive-throughs. The study included the following major sections:

I. Screeners
II. MaxDiff
III. Either CBC or ACBC
IV. Holdout Choice Tasks (4 tasks with 12 alternatives per task)

The issues we investigated included:

- Which random design strategy works better for CBC questionnaires, Complete Enumeration (minimal overlap) or Balanced Overlap (modest level overlap)?

- Can Adaptive CBC work reasonably for a 4-attribute study? Can it do as well as standard CBC?

- Would placing "Unacceptable" questions in priority over "Must-Have" questions improve Adaptive CBC results?

- Adaptive CBC offers individual-level (customized) utility constraints during HB estimation. Might that help or hurt holdout prediction?

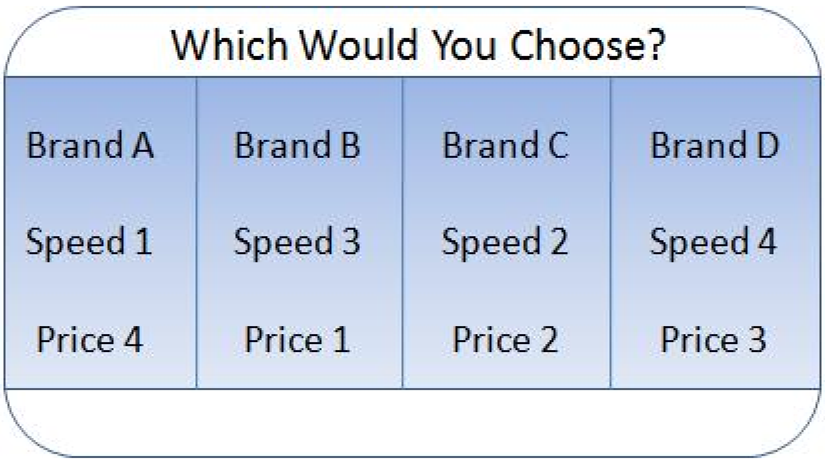
- Will respondents provide more consistent data and enjoy the questionnaire more if we give them a "consistency game"?

A more in-depth explanation of each of these questions and our conclusions regarding each follows.

Which random design strategy works better for CBC questionnaires, Complete Enumeration (minimal overlap) or Balanced Overlap (modest level overlap)?

First, let's provide some background about level overlap. In a recent newsletter, we showed the following example CBC task:

Figure 1



It has been common practice in CBC questionnaires to use minimal overlap when designing choice tasks. Minimal overlap simply means that we don't repeat a level within a task unless we have to. For example, consider a CBC study with four brands (A, B, C and D). We might design a choice task as shown in Figure 1.

In terms of statistical efficiency of main effects (the utility of each level considered independently), such tasks are optimal (assuming additivity and the logit rule). For this reason, our CBC software has used minimal overlap designs by default (the *Complete Enumeration* design method). Also, if each attribute has at most four levels, it has seemed natural to show just four products on the screen, as we get full coverage of the attribute list in each task and it limits the amount of information respondents have to evaluate at one time.

But, minimal overlap's allure of statistical efficiency and the desire to not overwhelm respondents with too many product concepts to consider per task has negative consequences that we recently have begun to appreciate.

It turns out that using these economical, minimal overlap designs encourages more simplification behavior and superficial information processing than the original card-sort conjoint approach. To illustrate this point, consider an extreme case: Imagine a respondent who has a "must-have" requirement that the product must be Brand B. Perhaps she works at the Brand B company, and therefore is intensely loyal. In each choice task, there is only one possible product she can choose. What are the outcomes? The respondent has an easy time answering the questionnaire (she simply scans each task for Brand B). The fit statistic from the individual-level HB model is extremely high since her answers are so predictable. And, we obtain a perfect hit rate for holdout tasks. But, we haven't learned anything about how she values the remaining attributes beyond brand. Yet, in a real product choice, there are multiple Brand B models for her to choose among that differ on performance and price. Our model might perform poorly in predicting her actual product choice.

Certainly, not all our respondents are so extreme. But, recent evidence suggests that perhaps a majority of respondents' behavior within CBC questionnaires can be explained assuming they are only reacting to at most two or three attribute levels (Gilbride and Allenby 2004, Johnson and Orme, 2007). To the degree that respondents establish a few must-have or must-avoid features, minimal overlap questionnaires are not very useful for developing much deeper insights at the individual level than those top-most requirements.

There is yet another benefit for level overlap that we've commented on before (Sawtooth Software 1998) but did not investigate here: having some level overlap significantly improves the precision of estimation of interaction effects.

Results

220 respondents completed a CBC questionnaire with 24 tasks, where each task had 5 concepts plus a None alternative (see the Appendix for an example choice task). There were four attributes in the study, including brand, service, wait time, and price for the meal. 107 respondents were randomly selected to receive a CBC questionnaire designed using *Balanced Overlap* and the other 113 respondents received a *Complete Enumeration* design. Both methods strive for a high degree of one-way and two-way level balance. But, Balanced Overlap permits a modest degree of level overlap (levels occasionally repeat within the same choice task).

Respondents completed 4 holdout choice tasks after the CBC tasks, where each task included 12 product alternatives (forced choice with no None, see the Appendix for an example task).

We use three measures of accuracy related to predicting the holdouts using part-worths estimated only from the 24 choice tasks:

- <u>Hit Rate</u>: If the respondent's sums of part-worths accurately predict which product concept the respondent chose in the holdout, it is counted as a "hit," otherwise it is counted as a "miss." We report the percent of hits across all respondents x 4 choice tasks.
- <u>Probability of the chosen holdout</u>: Using the logit rule, we can compute the likelihood that the respondent would choose the product the respondent actually did chose in the holdout. This is a continuous measure (from 0 to 1) that contains more information than the simple hit rate.
- <u>Share prediction accuracy (MAE—Mean Absolute Error)</u>: We use a market simulator to project the shares of choice for the subsample for the 12 product concepts in each holdout

task. These predicted shares are compared to the actual choice probabilities across the total sample (n=650). So, this analysis includes holdout respondents as well as holdout tasks, lessening the opportunity for overfitting to the holdouts. The absolute difference between predicted and actual shares is averaged across the 12 product concepts. Market simulations were conducted using the Randomized First Choice method, and the scale factor (exponent) was tuned to minimize MAE. With MAE, lower implies better prediction.

The results are given in Table 1:

Table 1
Performance of Balanced Overlap vs. Complete Enumeration

	Balanced Overlap (n=113)	Complete Enumeration (n=107)
Hit Rate	43.5%	41.5%
Probability of holdout choice	38.2%	35.9%
MAE	1.87	2.21

All three performance metrics were directionally in favor of the Balanced Overlap method, though the differences were not statistically significant. Had we used larger sample sizes, we might have been able to detect statistically significant differences.

This is the first study we are aware of that has compared the performance of Balanced Overlap to Complete Enumeration. Balanced Overlap also leads to more precise estimates of interaction effects than Complete Enumeration (Sawtooth Software 1998). Our analysis included just main effects, so we didn't try to exploit Complete Enumeration's weakness in that area. However, it could be argued that since our holdout tasks involved a great deal of overlap (showing 12 concepts on the screen required level overlap), this placed Complete Enumeration at a disadvantage. That may be true, but it is also a fact that most real-world choices involve level overlap. We'd prefer a questionnaire design technique that can lead to more accurate choices in those situations.

With the behavioral theory in favor of including some level overlap in choice tasks (encouraging extreme respondents to reveal deeper choice behavior), as well as our modest evidence here regarding potential performance gains for Balanced Overlap, we'd suggest researchers consider using Balanced Overlap instead of Complete Enumeration.

Can Adaptive CBC work reasonably for a 4-attribute study? Can it do as well as standard CBC?

We designed the Adaptive CBC (ACBC) system to be effective for studies with about 5 or more attributes. Our experience is that the traditional CBC approach should be hard to beat for situations involving about 4 attributes or fewer.

Included in the experiment we've been describing were some respondents who received ACBC questionnaires. In all previous research we'd conducted comparing ACBC to CBC, we had used at least 8 attributes and had always given ACBC a time advantage over CBC. In those studies, respondents spent from 50% more to triple the time completing the ACBC questionnaire than those with the CBC questionnaires. This time, we decided to give CBC every opportunity to do well, without the disadvantage of less time spent probing respondent preferences than ACBC. Respondents received 24 CBC tasks, where each task had 5 alternatives plus a None. The result was that the time to complete the CBC questionnaire was about one minute longer than completing the ACBC questionnaire.

As for the ACBC questionnaire, two of the four attributes had *a priori* known preference order, so it didn't make sense to ask about them in the BYO (configurator) question. We generated a pool of 18 product concepts based on the BYO-selected concept. We had to overcome some challenges to make a 4-attribute ACBC questionnaire work well. First, the ACBC v1 software we were using allowed researchers to generate near-neighbor product concepts by varying at minimum 1 attribute from the BYO concept. And, the software did not permit varying more than ½ of the attributes included in the BYO question when generating near-neighbor products. Thus, initially we were limited to varying at minimum 1 and at maximum 1 of the 2 BYO attributes when generating near-neighbor concepts. However, it became apparent to us that if one always retains one of two attributes at the BYO level while varying levels of the second attribute, this creates terrible designs from an efficiency standpoint.

We'll illustrate that point. With just 2 attributes included in the BYO section, imagine the respondent chooses levels 1 of both attributes in the BYO concept. If one of the attributes is always retained at the BYO-selected level, then there are only a limited number of combinations that can occur for attributes 1 and 2 in the near-neighbor concepts. In our design, the first attribute had 5 levels and the second attribute had 2 levels, so the prohibited combinations are marked with an X in the table below:

Table 2
Prohibitions Table, Given BYO Selection

		Attribute 2	
		Level 1	Level 2
Attribute 1	Level 1	X (BYO)	
	Level 2		X
	Level 3		X
	Level 4		X
	Level 5		X

Level 1 of attribute 1 can only show with level 2 of attribute 2. To avoid this problem, our programmers updated the ACBC design code to permit near-neighbor concepts to be generated that varied 0, 1, or 2 of the BYO-specified levels for the 2 attributes included in BYO[1].

[1] Having to customize the ACBC v1 software to do a 4-attribute design led us to provide greater flexibility in the design settings in an update to ACBC software, released the month following this paper's publication. Although it

Admittedly, changing *both* of the BYO levels might create a product that was less relevant for the respondent. But, if the respondent soon expresses that one of the BYO levels is a must-have, then the profiles where both levels are non-BYO levels would be dropped from further consideration (but included in the experimental design, as inferred non-choices), and new (more relevant) replacement concepts would be generated.

Time to complete. We included some "hidden variables" in our ACBC questionnaire, which we set to the server's system time, so we could compute elapsed time for different sections of the questionnaire. The timed section including the conjoint questions also included about 5 short paragraphs of explanatory text, as well as a 3-point ratings grid on 7 levels spanning the first two attributes (see Figure 2 further below). Our best guess is that the five paragraphs and the 7 ratings took about 1 minute to complete, so the times below could be reduced by about a minute to estimate the time actually spent in the conjoint sections.

ACBC 336 seconds (5.6 minutes)
CBC 406 seconds (6.8 minutes)

As stated earlier, we wanted to give CBC every opportunity to succeed, so we included 24 choice tasks at 5 alternatives per task. We wanted respondents to spend at least as much time in CBC as with the ACBC survey. Previous research by Johnson and Orme suggests respondents are quite reliable in CBC tasks through as many as 20 tasks (the authors didn't have data to investigate even longer questionnaires) (Johnson and Orme, 1996).

Table 3
Performance of ACBC vs. CBC

	ACBC (n=221)	CBC (n=220)
Hit Rate	42.5%	42.5%
Probability of holdout choice	34.0%	37.0%
MAE	2.02	2.04

Table 3 shows very similar results (no statistically significant differences) for ACBC and CBC in predicting the responses to the 4 holdout choice tasks. The good news is that both methods are quite robust. The null hit rate is 1/12 = 8.3%, so the hit rates achieved here are roughly 5x the chance level.

These results suggest no predictive edge for either ACBC or CBC for a 4-attribute study. Previous research has shown ACBC to have an edge for studies involving about 8 or more attributes, and also to be more interesting and engaging for respondents. The 24-task CBC survey was probably tedious for many respondents, but we didn't include any satisfaction questions within this study. Based on previous findings, we can only expect that respondents found the shorter ACBC questionnaire more interesting and less boring than the 24-task CBC.

generally isn't advisable, this study shows that there are occasions when no or all BYO-selected levels might be changed when generating new concepts to evaluate in the Screener section of ACBC.

Would placing "Unacceptable" questions in priority over "Must-Have" questions improve Adaptive CBC results?

Adaptive CBC (ACBC) is quite new, and there is much to learn regarding best practices. An early decision that Rich Johnson and I made (somewhat arbitrarily) was the position and order of the "Must-Haves" and "Unacceptables" questions. The suggested flow in the first release of the software (2009) was:

"Must-Haves" Priority Questionnaire Flow:
Screener #1
Screener #2
Screener #3
Must-Have #1
Screener #4
Must-Have #2
Unacceptable #1
Screener #5
Must-Have #3
Unacceptable #2
Screener #6
Etc.

When we were at the joint SKIM/Sawtooth Software European conference in Prague last May, our colleagues at SKIM presented a case study on ACBC. During that presentation, they commented that they felt that the Must-Have question allowed respondents too much power to eliminate levels (as unacceptable). Must-haves actually eliminate levels (as unacceptable) more aggressively than Unacceptables do.

For example, consider a fast-food restaurant survey in which for the first three screener questions, the respondent has marked that only McDonald's was a possibility. The Must-Have #1 question would ask that respondent to confirm that McDonald's was a Must-Have. If the respondent agreed, all other brands would be marked as Unacceptable.

However, if you use an Unacceptable #1 question in place of the Must-Have #1, the respondent would be asked if any of the non-McDonald's brands was Unacceptable. And, the respondent would only be able to indicate that *one* of these was unacceptable at that early point in the survey.

SKIM's argument made sense to us, so we decided to investigate this issue within this split-sample research project. The first version of the questionnaire had a layout as described above (Must-Have priority). The second version used this layout:

<u>"Unacceptables" Priority Questionnaire Flow:</u>
Screener #1
Screener #2
Screener #3
<u>Unacceptable #1</u>
Screener #4
<u>Unacceptable #2</u>
<u>Must-Have #1</u>
Screener #5
<u>Unacceptable #3</u>
<u>Must-Have #2</u>
Screener #6
Etc.

We found a modest increase in the percent of levels inferred/marked unacceptable if the Must-Have questions had priority (8.8% of levels marked unacceptable vs. 7.8%).

Table 4
Performance of "Unacceptable Priority" vs. "Must-Have Priority"

	Unacceptable Priority (n=110)	Must-Have Priority (n=111)
Hit Rate	43.7%	41.4%
Probability of holdout choice	34.7%	33.4%
MAE	1.68	2.35

All of these metrics (hit rates, probability of holdout choice, and share prediction accuracy) are directionally in favor of the version of the questionnaire where Unacceptables have priority (though the differences are not statistically significant).

Until more evidence is gathered, we suggest using the *Unacceptables Priority* questionnaire flow for ACBC, and we have changed the default behavior in the software to reflect this. (It would also be interesting to test a version of the questionnaire that drops the Must-Have question entirely, but we haven't tried that.) It is good news if this finding holds. It would mean that we can obtain even better results for ACBC with a very simple modification to the questionnaire flow.

Adaptive CBC offers individual-level (customized) utility constraints during HB estimation. Will that help or hurt holdout prediction?

A recent paper at the Sawtooth Software conference showed ways to incorporate rating or ranking information of levels within part-worth estimation of CBC data (Lattery 2009). The data could be treated as additional information, or as hard constraints.

ACBC offers constrained estimation under HB, using either global constraints (all respondents have the same preference order for levels within an attribute) or customized constraints (each respondent has a constrained order based on their idiosyncratic preferences as recorded in additional variables within the same SSI Web questionnaire).

Our empirical study involved 4 attributes. Two were unordered (chain brand, and location of order taker), and the other two had *a priori* preference order (wait time, and cost).

Prior to asking the BYO question, we asked respondents the following grid question:

Figure 2

We hoped that the 3-point scale offered here would allow respondents to express preference for some levels over others, without providing so much granularity that levels that were nearly tied in preference actually received different ratings. With the 3-point scale, many levels would be tied (that really didn't reflect large differences in preference for the respondent), and the ACBC exercise could further refine their relative preferences.

We estimated part-worths using all 4 attributes constrained, and a separate run using just the 2 *a priori* attributes constrained. Brand name of chain, and location of order taker involved customized constraints (using each respondent's answers to grid question directly above), and wait time and price involved global constraints.

The methodology for our utility constraints in HB follows the Simultaneous Tying approach first described by Johnson (2000). Johnson's results suggested that utility constraints should improve hit rates, but sometimes slightly degrade share prediction accuracy of simulators.

Under constrained estimation, we found inconsistent results (see Table 4).

Table 4
Unconstrained vs. Constrained Estimation for ACBC

	Unconstrained (n=221)	**4 Attributes Constrained (n=221)**	**2 Attributes Constrained (n=221)**
Hit Rate	42.5%	42.1%	43.4%
Probability of holdout choice	34.0%	33.2%	33.5%
MAE	2.02	2.07	2.05

Constraining part-worths for just 2 attributes (wait time and cost) based on global constraints performed directionally better on all measures than constraining all 4 attributes, though the differences were slight. Constraining 2 attributes performed directionally better than unconstrained on hit rate, but directionally worse on the other two measures. It is comforting to see that share prediction accuracy did not decline very much at all when using constraints.

We have some thoughts regarding why constraints didn't consistently improve hit rates for ACBC (as typically is seen with standard CBC), especially when including idiosyncratic constraints from the ratings grid:

1. ACBC contains relatively more information than standard CBC experiments, so the part-worths tend to have greater precision and have fewer reversals that need to be resolved by constrained estimation.
2. This study had only 4 attributes, with five or fewer levels per attribute. With larger and more typical ACBC studies, there will be less information per parameter to estimate and more opportunity for constraints to resolve out-of-order utility relationships and improve holdout hit rates.
3. The ratings grid was answered prior to the ACBC section. The holdouts followed the ACBC section. If the ratings grid were moved following the ACBC section, then respondents might have better resolved in their minds via the tradeoff exercise what levels they really preferred and what levels were of roughly equal preference. Asking the ratings grid following ACBC might lead to substantially better information for imposing utility constraints.

If we could repeat this study, we'd place the ratings grid (Figure 2) following the ACBC exercise.

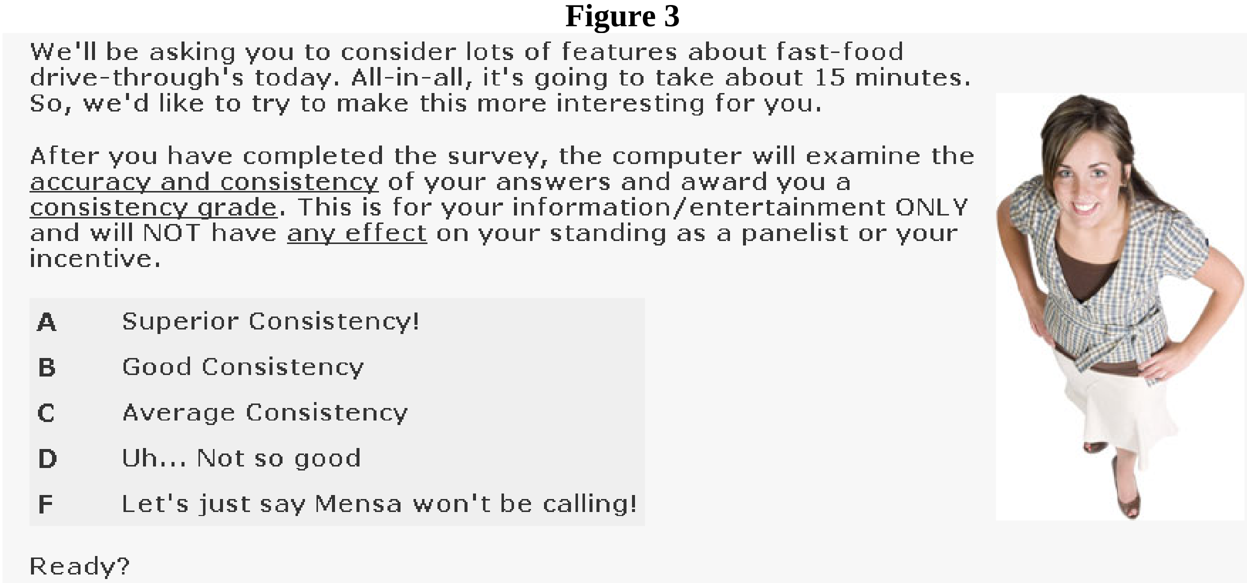
Will respondents provide more consistent data and enjoy the questionnaire more if we give them a "consistency game"?

Respondent attention and consistency is a common concern. Some respondents seem to speed or straightline, and panel management companies (including Western Wats, whom we used for this research) are vigilant to remove those who show a pattern of "bad" behavior over multiple studies. Researchers commonly throw out 10% to 20% of respondents, based on time to complete, straightlining, test-retest consistency checks, and low fit statistics from conjoint or MaxDiff.

We wondered whether we could induce respondents via a "consistency game" to provide more consistent data, slow down, and pay a bit more attention, hopefully without introducing bias. In our split-sample experiments, half of respondents received a consistency game and the other half did not.

The game went as follows, early on in the interview, respondents in the consistency game were told:

Figure 3


We'll be asking you to consider lots of features about fast-food drive-through's today. All-in-all, it's going to take about 15 minutes. So, we'd like to try to make this more interesting for you.

After you have completed the survey, the computer will examine the accuracy and consistency of your answers and award you a consistency grade. This is for your information/entertainment ONLY and will NOT have any effect on your standing as a panelist or your incentive.

A	Superior Consistency!
B	Good Consistency
C	Average Consistency
D	Uh... Not so good
F	Let's just say Mensa won't be calling!

Ready?

Later, just prior to the holdout choice tasks, we reminded them…

Figure 4

After you have made these last four selections, the computer will evaluate the accuracy and consistency of the answers you've given to this survey and award you a consistency grade.

This is for your information/entertainment ONLY and will NOT have any effect on your standing as a panelist or your incentive.

A	Superior Consistency!
B	Good Consistency
C	Average Consistency
D	Uh... Not so good
F	Let's just say Mensa won't be calling!

Ready?

We also asked respondents an open-end question as follows (after the 2nd holdout task was completed):

Figure 5

After a couple more questions, the computer will grade your consistency...

How do you feel about the Consistency Grading idea we've introduced in this survey? Fun? So-So? Annoying?

(Please type a few words, or a sentence...)

The consistency grading was done based on responses to the 4 holdout choice tasks. The last two were repeats of the first two, so we could assign a consistency grade based on number of holdouts (see Appendix for example) answered consistently (0=C, 1=B, 2=A; like many graduate school professors, we didn't award any D's or F's). We hoped that respondents would slow down and be more consistent, and that they might also find the game interesting.

It turned out that the game didn't seem to cause respondents to slow down any, but there probably was a modest gain in the fit statistic for MaxDiff (RLH*1000) due to the game. Respondents receiving the consistency game had a fit of 589 vs. 561 for those not receiving the game (t=2.1). We also looked at the test-retest reliability of holdouts (since the last 2 holdouts were repeats of the first two holdouts, with order of concepts scrambled). The test-retest rate was directionally better for the consistency game group, but it was not a statistically significant difference.

An examination of the open-ends revealed that 41% of respondents liked the game, 42% were indifferent, and 17% disliked the game.

So, we conclude that a consistency game had very little effect on respondents' performance, and it risks annoying 1/5 of the respondents. We don't think our particular consistency game worked very well and do not recommend it.

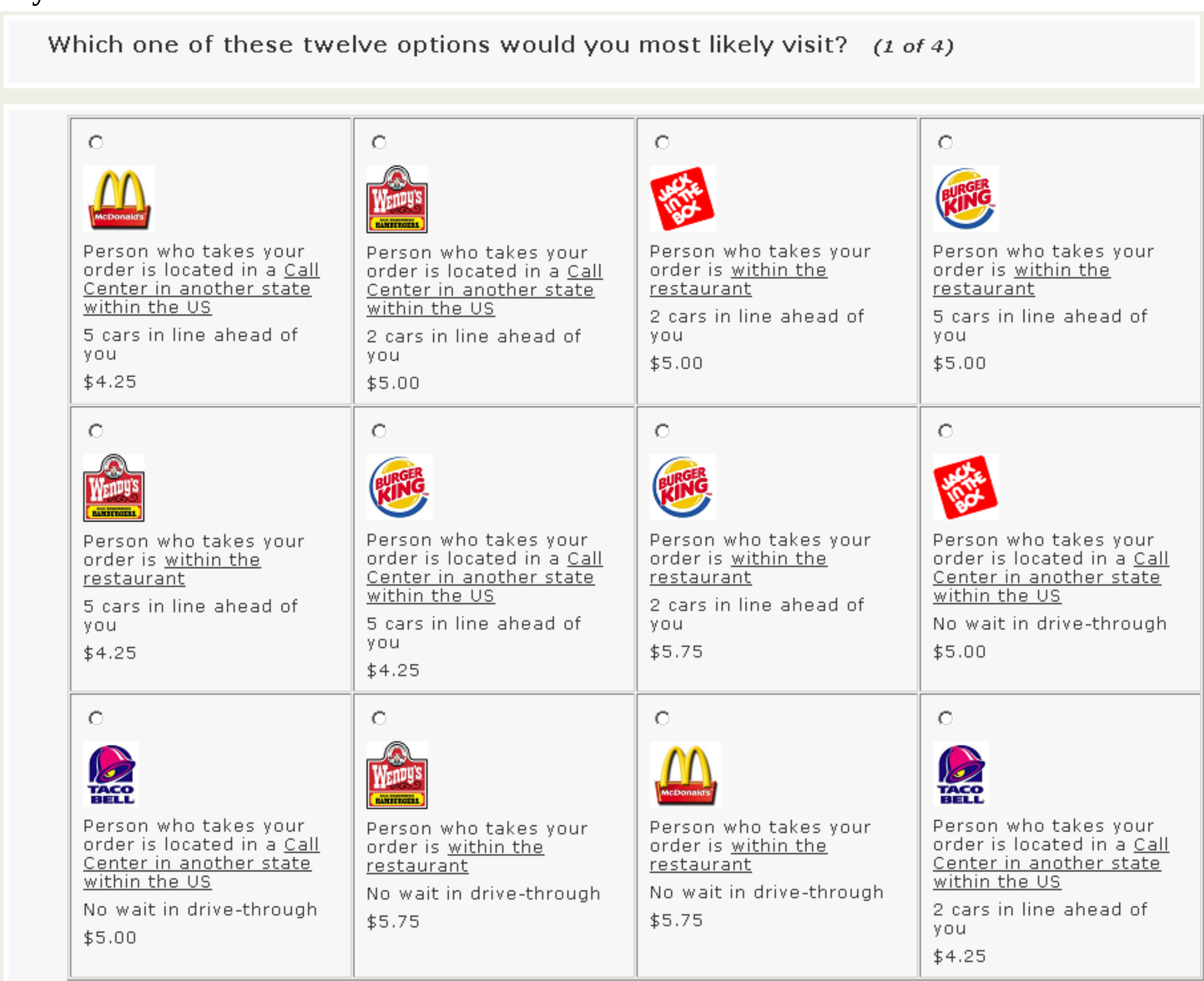
Appendix

Layout of CBC Task:

If these were your only options, which would you choose?
Choose by clicking one of the buttons below:

Taco Bell	Wendy's	Burger King	Jack in the Box	McDonald's	
Person who takes your order is within the restaurant	Person who takes your order is located in a Call Center in another state within the US	Person who takes your order is located in a Call Center in another state within the US	Person who takes your order is within the restaurant	Person who takes your order is within the restaurant	NONE: I wouldn't choose any of these.
2 cars in line ahead of you	2 cars in line ahead of you	5 cars in line ahead of you	No wait in drive-through	5 cars in line ahead of you	
$5.75	$5.00	$5.75	$4.25	$5.00	

Layout of Holdout Task:



Which one of these twelve options would you most likely visit? *(1 of 4)*

McDonald's Person who takes your order is located in a Call Center in another state within the US 5 cars in line ahead of you $4.25	Wendy's Person who takes your order is located in a Call Center in another state within the US 2 cars in line ahead of you $5.00	Jack in the Box Person who takes your order is within the restaurant 2 cars in line ahead of you $5.00	Burger King Person who takes your order is within the restaurant 5 cars in line ahead of you $5.00
Wendy's Person who takes your order is within the restaurant 5 cars in line ahead of you $4.25	Burger King Person who takes your order is located in a Call Center in another state within the US 5 cars in line ahead of you $4.25	Burger King Person who takes your order is within the restaurant 2 cars in line ahead of you $5.75	Jack in the Box Person who takes your order is located in a Call Center in another state within the US No wait in drive-through $5.00
Taco Bell Person who takes your order is located in a Call Center in another state within the US No wait in drive-through $5.00	Wendy's Person who takes your order is within the restaurant No wait in drive-through $5.75	McDonald's Person who takes your order is within the restaurant No wait in drive-through $5.75	Taco Bell Person who takes your order is located in a Call Center in another state within the US 2 cars in line ahead of you $4.25

References

Gilbride, Timothy and Greg M. Allenby (2004), "A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," *Marketing Science*, 23, 3, (Summer) 391-406.

Johnson, Rich, (2000), "Monotonicity Constraints in Choice-Based Conjoint with Hierarchical Bayes (2000)," Technical Paper available at www.sawtoothsoftware.com.

Johnson, Rich, and Bryan Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint Studies?" American Marketing Association's ART/Forum, available at www.sawtoothsoftware.com.

Johnson, Rich and Bryan Orme (2007), "A New Approach to Adaptive CBC," Sawtooth Software Conference Proceedings, Sequim, WA.

Lattery, Kevin (2009), "Coupling Stated Preferences with Conjoint Tasks to Better Estimate Individual-Level Utilities," Sawtooth Software Conference, Sequim, WA, forthcoming.

Sawtooth Software (1998), "On Interaction Effects and CBC Designs," Sawtooth Solutions, Sequim, WA. Available at www.sawtoothsoftware.com/education/ss/ss8.shtml.

Sawtooth Software (2008), "Update on Conjoint Usage Survey," Sawtooth Solutions, Sequim, WA.