Sawtooth Software

RESEARCH PAPER SERIES

Including Holdout Choice Tasks in Conjoint Studies

Bryan Orme
Sawtooth Software, Inc.

1457 E 840 N   Orem, Utah
+1 801 477 4700
www.sawtoothsoftware.com

Including Holdout Choice Tasks in Conjoint Studies

Bryan Orme, Sawtooth Software

Note: this work represents an update to our 1997 and 2010 technical papers of this same title.

What Is a Fixed (Holdout) CBC Task?

For many years now, our CBC software's user interface automatically has "suggested" a default two fixed (holdout) choice tasks be included in CBC surveys (the user can of course change the default). Fixed (holdout) choice tasks look to the respondent just like the other experimentally designed ("Random[1]") CBC tasks. What makes them different is that the attribute level combinations in these tasks are fixed—held the same—across respondents. Importantly, these fixed choice tasks are usually held out from utility estimation.

The researcher specifies exactly which combinations of attribute levels to show in each fixed choice task. For example, the client's *base case* scenario (a starting point for later market simulations) may be included as one of the fixed holdout tasks. The part-worth utilities estimated from the responses to the experimentally designed tasks (the "Random" tasks) are used in a market simulator to predict respondents' choices to the holdout tasks. The hope, of course, is that the predictions closely resemble the answers to the fixed holdout tasks. A poor match could be an indication that something went wrong in data collection, data coding, utility estimation, market simulation, or that the respondents answered very inconsistently.

Are Two Holdout Tasks Enough?

Although our CBC software suggests a default two fixed CBC tasks, *this does not mean that two is enough*! We wanted to encourage researchers to put holdout tasks within CBC surveys, so we programmed the software by default to insert at least a couple holdout tasks. But, we didn't intend to imply that two was enough choice tasks for every situation or even for most situations.

Two holdout choice tasks are probably only enough to reveal if you've made a *major* error in data collection, data processing, and analysis. Being able to show that the market simulator can predict the aggregate choices fairly well for two holdout choice tasks gives you and your clients additional confidence that the CBC model is working. This may be especially helpful if you or your client is somewhat new to conjoint analysis.

If you plan to use holdout choice tasks to make judgments about which model specification works

[1] Even though Sawtooth Software's experimentally designed CBC tasks are often called "Random," they are very carefully chosen to satisfy level balance and orthogonality. They have been called "Random" due to the fact that respondents are randomly selected to receive one of many available versions (blocks) of the experimentally designed choice tasks.

better (e.g. main effects vs. a model that also includes certain interaction terms) or to compare the predictive validity of one conjoint method to another, then *just two holdout tasks does not provide enough information*. A recent white paper Keith Chrzan (Chrzan 2015) suggests that 5 or so holdout choice tasks are often enough to indentify the better of two models reliably if the different models represent moderate to large differences in preference functions; but perhaps 15 or more holdouts are required if the differences involved are relatively small.

Reasons Holdout Tasks Are Useful

- They provide a proximal indication of face validity, measured by the utilities' ability to predict choices not used in their estimation. This can help expose errors in survey programming, data processing, or fielding, as well as help clients gain additional confidence that the conjoint utilities are useful for predicting complex choices.

- They provide a check on the scaling of the utilities. CBC utilities should already have appropriate scale to predict CBC-looking holdout tasks for the same respondents, but ratings-based conjoint and ACBC typically require scale parameter adjustment to predict CBC-looking holdouts well. If the most popular concepts are over-predicted, then the scale parameter (Exponent, in Sawtooth Software's simulator) should be reduced. If the predictions are too flat, then the scale parameter should be increased.

- They permit identification and removal of inconsistent respondents.

- They can be used for testing specific product configurations under consideration. Much value can be added by directly assessing respondent reaction to these concepts.

Designing Holdout Choice Tasks

It's hard to design good holdout concepts without some prior idea of respondent preferences. There's no point in asking people to choose among concepts where one dominates in the sense that everyone agrees which is best. Similarly, it's good to avoid presenting concepts that are equally attractive, since equal shares of preference would be predicted by a completely random simulator. If you present triples of concepts, it's probably best if their shares of choices are somewhere in the neighborhood of 50/30/20.

Over the last 25 years, most CBC studies and holdout questions have featured *minimal overlap*, which means that an attribute level is not repeated within a choice task (unless there are more products to show in the choice task than levels in the attribute). Over the last few years, the minimal overlap questionnaire has been shown to be less effective at estimating preferences for respondents who do not trade off attribute levels, but use heuristic decision rules such as *must haves, unacceptables,* or other non-compensatory decision strategies. For example, if a respondent requires a certain brand, and there is only one such brand offered per choice task *and* holdout task, it makes it trivial for this respondent to answer the CBC questionnaire, and trivial for the estimated utilities to predict this person's holdout choice. High fit to the data in this case does not necessarily indicate that we've been successful at estimating this respondent's complex preferences beyond brand requirement, and we may find it quite difficult to predict this respondent's real world choices, where the required brand typically has multiple variants on the other dimensions.

Since most respondents probably use non-compensatory rules when facing complex conjoint questionnaires, there is a strong movement toward using CBC questionnaires featuring level overlap (such as provided using CBC software's *Balanced Overlap* design approach and the approach featured in our Discover-CBC online platform). We strongly recommend using holdouts that feature level overlap, because real world market decisions also tend to involve multiple product offerings featuring level overlap. Also, it is important to use a healthy amount of level overlap on attributes that respondents are known to screen on (apply as non-compensatory rules).

When conducting CBC studies, if you plan to do segmentation with latent class analysis, it's wise to consider the kinds of groups you expect to get, and to design products in holdout choice sets so that one alternative will be much more preferred by each group. This maximizes your ability to confirm the validity of the multi-group Latent Class simulator.

It isn't necessary to have very many holdout sets to check the face validity of your utilities or to rescale ratings-based conjoint data to reasonably predict choice probabilities. However, if you want to use those choices to identify and eliminate inconsistent respondents, you need several choice sets.

The position of the holdout tasks is important. It is well known that the first CBC tasks contain the most noise and the lowest scale. Respondents tend to learn through the process of completing multiple CBC tasks. They tend to rely more on brand in the first few tasks and more on price in later tasks, and the use of the None choice increases in later tasks. We generally suggest that holdout tasks be spaced evenly throughout a CBC questionnaire (e.g. in the 4th, 8th, and 12th positions if using a 16-task survey).

For CBC, ACBC, CVA, and ACA studies, CBC-looking holdout tasks can be included in the computer-administered interview using the SSI Web System. Here is an example of a holdout choice task:

If you were shopping for a credit card and these were your only options, which would you choose?

Visa	MasterCard	Discover	Visa
No Annual Fee	$30 Annual Fee	$60 Annual Fee	$30 Annual Fee
15% Interest Rate	12% Interest Rate	9% Interest Rate	9% Interest Rate
Frequent Flier Program	No Frequent Flier Program	Frequent Flier Program	No Frequent Flier Program
$4,000 Credit Line	$6,000 Credit Line	$2,000 Credit Line	$4,000 Credit Line

It is not very useful to include a "None" option in holdout choice tasks when these are paired with traditional conjoint exercises which don't have a "None" option. This would make it difficult to compare the results of market simulations to holdout choices.

Finally, if you do have several holdout choice sets, it's useful to repeat at least one of them so you can obtain a measure of the reliability of the holdout choices. Suppose your conjoint utilities are able to predict only 50% of the respondents' holdout choices. Lacking data about reliability, you might conclude that the conjoint exercise had been a failure. But if you were to learn that repeat[2] holdout tasks had reliability of only 50%, you might conclude that the conjoint utilities were doing about as well as they possibly could, and that the problem lies in the reliability of the holdout judgments themselves.

Should We Use Holdout Choice Tasks in Utility Estimation?

At first glance, this question seems illogical, since holdout tasks are by definition held out of utility estimation. However, after we have used the holdout tasks for their purpose (to check the face validity of the data, to identify bad respondents, etc.) it would seem a waste to throw away the responses that have been paid for with respondent time and client money. If the experimentally designed tasks are already efficient for estimating the parameters of interest, from a statistical standpoint additional holdout tasks should only provide more information for improving the utility estimates. Even so, this opens the possibility of specific psychological context effects (owing to the fixed tasks seen by all respondents) affecting the part-worth utility estimates, so we offer this suggestion with a caution.

[2] When repeating holdout tasks to assess test-retest reliability, we'd recommend rotating the concept order so that respondents who are straightlining the CBC questions don't achieve 100% test-retest reliability.

An interesting prospect to consider is the value of including the client's base case market simulation scenario as a fixed task within a CBC project, to be included in utility estimation. If the data analysis plan involves testing various what-if simulations around a particular base case, then it would seem beneficial to include that base case itself as one of the tasks used in utility estimation.

Within-Sample or Out-of-Sample Holdouts?

The most rigorous way to test the predictive validity of marketing analytics models is out-of-sample. Out-of-sample validation involves using the part-worth utilities from one group of respondents (the *calibration* respondents) to predict a different set of respondent's choices (the *holdout* respondents), where the two groups of respondents are selected to as closely match as possible on variables that affect choice. Actual market purchases are another example of out-of-sample data.

Out-of-sample validation guards against overfitting. Overfitting is the situation when parameters included in the model provide better fit to the data used in building the model, but actually make predictions worse for new data. Another form of overfitting is when individual-level models that seek to explain heterogeneity (individual respondent tastes) provide excellent fit to within-sample choices, but work less well in predicting new respondents' choices. There is ongoing research and debate about the best way to capture heterogeneity in conjoint data while providing enhanced out-of-sample predictive validity.

Unfortunately, few commercial conjoint analysis studies have the luxury of doubling the sample size to collect out-of-sample holdout data that will only be used to check the validity of the calibration sample! Robust conjoint studies for academic purposes (publishing in conferences or journals) should justify the expense of out-of-sample data, but it may seem hard to swallow otherwise. Here is a possible way to overcome this hurdle:

- Imagine a situation in which 600 respondents will receive a CBC questionnaire with 12 tasks.
- The researcher adds three holdout concepts in the questionnaire, in positions 4, 7, and 10. The other nine tasks are experimentally designed ("Random" tasks) for the purposes of estimating the part-worth utilities.
- From Sawtooth Software's CBC system, the researcher exports the experimental design to a .CSV file and modifies the fixed tasks so that they rotate across versions of the questionnaire. Four blocks of 150 respondents each receive one of four versions of the three holdout tasks, leading to a total of 12 unique holdout tasks for validation.
- The researcher estimates part-worth utilities for the sample (n=600) using the nine experimentally designed tasks. These are used within a market simulator to predict the choice likelihood for the alternatives in the 12 holdout tasks (n=150 choices per holdout task).
- After the holdout tasks serve their purpose (e.g. check face validity of the data, guide

model specification and utility estimation method), they are put back into utility estimation for developing the final set of part-worths to be delivered to the client. Each respondent now has 12 tasks for utility estimation.

Note that the above suggestion is a practical approach to balance the demands of practitioner study budgets and achieve a good degree of out-of-sample academic rigor. The client doesn’t feel that the holdout data were wasted and the respondents weren’t compelled to answer an overly long CBC questionnaire.

We admit, however, that the out-of-sample data are not entirely out-of-sample in this illustration: 600 respondents’ part-worth utilities are being used to estimate the holdout responses of 150 respondents, all of whom come from the 600 respondents. If this concerns you, with some extra work you can conduct a multi-step jackknife procedure, where 450 respondents predict the truly out-of-sample 150 respondents’ data four separate times until all out-of-sample data have been predicted.

References:

Chrzan, Keith, “How Many Holdouts for Model Validation,” Sawtooth Software Research Paper Series, available at: http://www.sawtoothsoftware.com/support/technical-papers.