Sawtooth Software

RESEARCH PAPER SERIES

How Many Holdout Tasks for Model Validation?

Keith Chrzan
Sawtooth Software, Inc.

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

How Many Holdouts for Model Validation?

Keith Chrzan, Sawtooth Software
February 2015

Background

Analysts sometimes add holdout questions to their conjoint surveys to test the way they have specified their models. For example, we may want to make sure that we have the correct mix of part-worth and linear functions, say, or of main effects and interactions. Using holdouts can be a kind of insurance that allows us to verify that the modeling choices we make for our sample of respondents will hold up outside of our study.

Two uncertainties face analysts wanting to use holdouts for model validation. First, holdouts may be "in sample" or "out-of-sample." If each respondent completes 12 choice-based conjoint (CBC) questions to be used to build the statistical model and another three CBC questions that are "held out" for testing the model, these are in-sample holdouts. With out-of-sample holdouts, on the other hand, one set of respondents answers the CBC questions used for utility estimation and a different set of respondents answers a different set of CBC questions used for validity checking. Controversy exists regarding whether one needs out-of-sample holdouts or whether in-sample holdouts will suffice.

The second uncertainty concerns how many holdouts we need to ask to validate our model specification. SSI Web sets the default number of holdout questions at two, indicating that Sawtooth Software recommends the use of holdouts, but up until now there has been very little evidence about how many holdouts might be needed.

In-sample Versus Out-of-sample

This controversy largely depends on one's objectives: in-sample holdouts allow us to see how well our model generalizes to the broader universe of possible CBC *questions* for a given sample of choosers while out-of-sample holdouts allows us to assess how well our sample of respondents and questions generalizes to the broader universe of *choosers AND choice questions*.

The remainder of this white paper reports the results of an experiment which used carefully crafted artificial respondents to address how many holdout questions its take to distinguish a true underlying model from a false one.

Data Creation

It's easy to fool yourself when using artificial respondents, because no matter how closely you try to mimic the choices of human respondents, artificial respondents really aren't human: everything they do depends on what you program into them. That said, the statistical model we use for our choice-based conjoint studies has some pretty strong assumptions about respondent behavior and how respondent preferences map onto choices, which we can leverage to make our

artificial respondents more realistic; in particular they include a substantial role for randomness or response error in the decision process. Moreover, we can build our artificial respondents' utilities on a foundation of estimated utilities from real human respondents; this preserves realistic patterns of respondent heterogeneity and of correlations among utilities.

In this study we take respondent utilities from an R&D study of 620 respondents. To create effects for our test of holdouts to detect, we modify the utility data to create both a non-linear effect and an interaction effect. Finally, we add a realistic amount of random response error (an amount consistent with the kind of test-retest hit rates we see in commercial conjoint data) and we will have as realistic of artificial respondents as can be built.

The R&D study's 620 respondents completed a CBC study with two each of five-, four- and three-level attributes. Randomly dividing the artificial respondents in half, 310 complete 10 estimation CBC questions and 20 in-sample holdout CBC questions; the other 310 complete only the 20 holdout questions and will serve as our out-of-sample holdouts. Importantly, the holdout questions feature a completely random construction – in holdout questions we want lots of level overlap or else they are too easy to predict, especially for in-sample analyses (Orme 2014).

Typically we test a model's ability to discern different sizes of effects. Thus three versions of the validation analysis vary the size of the non-linear and interaction effects whose detectability we want to measure. The (small/medium/large) effect sizes for the mean utilities are shown below in parentheses. All other utilities remain as they were in the R&D study.

Table 1
Mean Part-Worth Utilities for Artificial Respondents

Attribute Level	Part-worth From Empirical Study	Part-worth With Effects
Google Nexus	-0.68	-0.68
Amazon Kindle Fire	-0.19	-0.19
Apple iPad	1.26	1.26
Samsung Galaxy	0.3	0.3
Microsoft Surface	-0.68	-0.68
7 inch	-0.6	-0.6
8 inch	-0.17	-0.17
9 inch	0.3	0.3
10 inch	0.46	0.46
16 GB	-1.41	-1.41
32 GB	-0.3	-0.3
64 GB	0.4	0.4
128 GB	1.31	(2.31/3.31/4.31)
GB1H1	-1.45	(-0.45/0.55/1.55)
GB1H2	-1.24	-1.24
GB1H3	-1.08	-1.08
GB2H1	-0.11	-0.11
GB2H2	0.1	0.1
GB2H3	0.27	0.27
GB3H1	0.97	0.97
GB3H2	1.19	1.19
GB3H3	1.35	1.35
$169	5.95	5.95
$199	1.1	1.1
$299	-0.74	-0.74
399	-2.58	-2.58
$499	-4.42	-4.42

So the "small effect size" condition adds a constant 1.0 to the utility of each the 128 GB level of memory and to the interaction of the first levels of GB and height for each individual respondent. Medium and large effect sizes add 2.0 and 3.0, respectively, to the values in the empirical study, again uniformly across respondents, with no variation, making these solid effects that should be visible.

Analyses

We fit two HB models to each of the effect size data sets, one for the model we know to be true (with the non-linear effect for GB and with the interaction of GB and height) and one for a misspecified model (with main-effects only model and a linear effect for GB). For each of the three data sets, simulating the holdout choices with the resulting utilities allows us to measure the fit of both sets of predictions to the actual holdout choices.

Often analysts look at the holdout tasks and simply check whether one model fits them more often than another – call this looking at the number of correct predictions. More rigorously, an analyst might want to know whether a difference in two models' predictions is statistically significant.

Results – number of correct predictions

For some holdout choice sets the correctly specified model predicts better and for others the misspecified model predicts better. In fact, for the large effect model the correctly specified model fits about 68% of the holdouts better while 32% fit better with the misspecified model. For the medium effect model, 65% of holdouts favor the correctly specified model and 35% the misspecified model. For the small effect model the correctly specified and misspecified models split the number of holdouts evenly. These proportions were similar for in-sample and out-of-sample holdouts, so the analysis below applies to both.

Clearly an unlucky selection of holdouts can validate the wrong model. On the other hand, when a given holdout has over a 60% chance of pointing to the correct model, it will not be too difficult to identify the correct model – with confidence rising as the number of holdouts rise. Because the results depend on the particular set of holdouts used, the table below reports the results of bootstrap sampling from the population of 20 holdouts included in the study. Numbers in the body of Tables 2 and 3 show how often the holdout hit rate for the correctly specified model *directionally* exceeds that for the misspecified model in bootstrap samples.

Table 2
Percent of Directional Wins for True Model – In Sample

	In Sample Holdouts							
Effect Size	**2**	**3**	**4**	**5**	**7**	**10**	**15**	**20**
Small (1)	50	52	52	57	62	62	63	68
Medium (2)	76	80	80	87	93	96	97	97
Large (3)	90	91	96	97	98	100	100	100

For the medium and large effects conditions, with even two holdout questions we identify the correct model over three quarters of the time and additional holdouts improve our chances further: by four holdouts and medium effect sizes we identify the correct model 80% of the time and we achieve 97% confidence by 15 holdouts.

Shifting to out-of-sample holdouts, and using mean squared error (MSE) as our measure of fit, the bootstrap sampling analysis results appear in Table 3. Two holdout questions identify the correct model over 70% of the time. Note that the likelihood of identifying the correct model is lower for out-of-sample holdouts, but still very high.

Table 3
Percent Directional Wins for True Model - OOS

	Out-of-Sample Holdouts							
Effect Size	**2**	**3**	**4**	**5**	**7**	**10**	**15**	**20**
Small (1)	*Misspecified*		*model wins as often as properly specified model*					
Medium (2)	71	76	80	80	86	87	97	100
Large (3)	74	79	82	83	91	94	97	100

For the medium and large effects conditions, with even two holdout questions we identify the correct model over half the time and additional holdouts improve our chances further: by four holdouts we identify the correct model 80% of the time and we exceed 95% confidence by 15 holdouts.

Because in the small effects scenario the correctly specified model and the misspecified model win equally often among the 20 holdout questions, and by very nearly identical amounts in terms of hit rates, increasing the number of holdouts does not help in the small effect case for the out-of-sample holdouts. Had we looked at a larger universe of holdout questions we would eventually have found a slight advantage for the correctly specified model. Even so, the number of holdouts one would need to ask to confidently identify the correct model would be prohibitive (and a poor value for the investment since the effect on simulation accuracy is so slight).

Results – statistical significance

The analysis above applies if one looks for a directional win in terms of hit rate or MSE. Of course for any pair of models, one of them will nearly always fit holdout data better than the other. So the fact that one model fits better than another does not tell us whether that model would fit the holdout data better had we used a slightly different sample of respondents on a different set of fielding dates. For this we need to know if one model fits the holdout data *significantly* better than another. Thus some analysts prefer to take things a step further and conduct statistical testing on holdouts to distinguish between the correctly specified and misspecified models. For in-sample holdouts one would use a dependent t-test for means, comparing the respondent-level hit rates for the true and misspecified models. The test for out-of-sample predictions is more complex. The MSE does not lend itself to statistical testing so we use a test of a difference in dependent correlations (Cohen and Cohen 1983). It's a bit of a nasty equation but send me an email and I'll be happy to provide you with an Excel spreadsheet containing the formula.

Tables 4 and 5 show the percentage of times the correct model *significantly* (at $p < 0.05$) beats the misspecified model by number of holdouts and by effect size.

Table 4
Percent of Significant Wins for True Model – In Sample

	In-Sample Holdouts							
Effect Size	**2**	**3**	**4**	**5**	**7**	**10**	**15**	**20**
Small (1)	*Misspecified model wins as often as properly specified model*							
Medium (2)	15	21	25	31	35	42	51	55
Large (3)	37	51	61	63	75	76	94	96

Notice that for medium-sized effects the out-of-sample holdouts give us more power to distinguish the true from misspecified model than do the in-sample holdouts. Small effects are so difficult to discern and large ones so easy that both types of holdouts are about equally sensitive.

Table 5
Percent of Significant Wins for True Model – OOS

	Out-of-Sample Holdouts							
Effect Size	**2**	**3**	**4**	**5**	**7**	**10**	**15**	**20**
Small (1)	*Misspecified model wins as often as properly specified model*							
Medium (2)	16	21	37	44	46	59	88	92
Large (3)	25	46	47	65	70	81	95	99

Conclusions

In this study no affordable number of holdout questions allows us to identify reliably correct from misspecified models in the face of small effect sizes.

With medium and large effect sizes it takes a few more than two holdouts to validate the specification of our CBC models (at levels nearing levels of confidence used in statistical testing), but not necessarily a lot more.

We would expect the results of a study like this one to depend a great deal on the details. For example, models with more numerous medium size effects might be easier to discriminate among than the medium effect model in this study, with its single interaction and its single non-linearity. On the other hand, the effects that would have managerial significance might well be smaller than the medium and large effects used in this study, making the differences harder to detect. Or again, perhaps we know in advance which effects to examine, in which case we can tailor holdouts to reveal them.

Still, analysts may take comfort in the fact that a handful of randomly constructed holdouts (not too many more than the SSI Web default setting of two) should be enough to give them a high level of confidence that they can avoid large or medium size errors in model specification, though errors smaller than 2.0 utility points will require larger numbers of holdouts to detect. Moreover, with a larger handful of holdouts, analysts can identify *statistically significant* differences, at least when those differences are medium to large.

Both in-sample and out-of-sample holdouts seem able to do this work for us. In-sample holdouts allow validation to benefit from over-fitting of our models (i.e. fitting them to include idiosyncratic aspects of the respondents or the context that would not generalize well outside the survey). Thus the primary driver for deciding between in-sample and out-of-sample holdouts should perhaps be the kind of generalizability one seeks.

References

Cohen, J. and P. Cohen (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 2nd ed.* Hillsdale: Lawrence Erlbaum.

Orme, B. (2014) "Including Holdout Choice Tasks in Conjoint Studies," available at http://www.sawtoothsoftware.com/support/technical-papers/general-conjoint-analysis/including-holdout-choice-tasks-in-conjoint-studies-2014.