Sawtooth Software

RESEARCH PAPER SERIES

Diagnostics for Random Respondents in Choice Experiments

Keith Chrzan and Cameron Halversen
Sawtooth Software, Inc.

© Copyright 2020, Sawtooth Software, Inc.
3210 N. Canyon Rd., Provo, Utah
+1 801 477 4700
www.sawtoothsoftware.com

Diagnostics for Random Respondents in Choice Experiments

Keith Chrzan and Cameron Halversen, Sawtooth Software, Inc.

September 2020

1.0 Background

Researchers suspect that some proportion of their survey respondents are responding randomly, and they use several methods to try to identify these respondents. Survey researchers can measure survey length to identify speeders, they can look at response patterns to identify straightliners, they can build consistency traps into their surveys to catch careless respondents, etc. Survey-based choice experiments feature sometimes-difficult repeated questions that use the same format over and over, which may be even more subject to respondent inattention than simpler, more direct, questions. Surely it's less cognitively challenging for a respondent to report her nationality than it is for her to answer a battery of dozen or more questions, in each of which she identifies her favorite among three product profiles that change from question to question on each of several dimensions.

Some standard procedures can catch respondents who use patterned responses to complete their surveys quickly, for example in an unlabeled CBC where a respondent who always chooses the 3rd alternative in all 10 choice questions, something that we would expect to happen for fewer than two in 100,000 conscientious and truthful respondents. Other standard measures don't work as well, however: is someone who spent only two seconds per question a speeder who should be deleted, or is he someone who cares only about a single attribute and who can express his honest preferences rapidly, without processing the entire list of attributes?

We want to restrict our focus to random responders, and to three methods that have been proposed to identify them in stated choice experiments like choice-based conjoint and maximum difference (MaxDiff) scaling. We seek to compare these three methods to see which one performs best in identifying random responders.

2.0 Three Methods for Identifying Random Respondents in Choice Experiments

Orme (2019) recommends creating a large number, say 1,000, artificial respondents, programmed to answer the choice experiment's questions randomly. Hierarchical Bayesian (HB) multinomial logit (MNL) of this response data will produce a random set of utilities, and, more interestingly, a fit statistic called root-likelihood, or RLH. RLH is higher for respondents whose utility model fits their observed choices well and lower for respondents whose utility

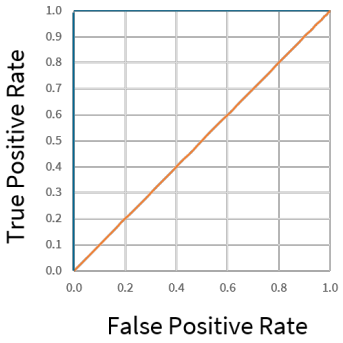
model fits their choices poorly. Random responders will have models that tend to fit their (random, unpredictable) choices poorly, so from this set of random responders, we can identify a cutoff RLH below which we expect some percentage (e.g. 95%) of random responders to fall. We can then apply this cutoff to our survey data to identify likely random responders.

Two methods employ latent class (LC) MNL to identify segments of random choosers. In the first of these, Hoogerbrugge and de Jong (2019) devise a clever method that again uses artificial random respondents. Again, we generate a large number of random respondents, e.g., about a sixth as many artificial random responders as there are survey respondents. We combine the response data from these random responders with our survey response data and we run a LC-MNL with a large number of classes, say 20. We then look for classes comprised mostly of our artificial random respondents and we classify the survey respondents in those classes as random responders.

The other LC method is a little easier in that it doesn't involve the creation of any artificial random responders. A method called scale-adjusted latent class, or SALC (Magidson and Vermunt 2007) can estimate both (a) latent "preference" classes or segments (i.e. segments with different tastes and preferences about the attribute levels) and (b) latent "scale" classes (i.e. segments of respondents with different amounts of response error, reflected in the logit scale parameter). Constraining one of the scale classes to have scale equal to zero will identify respondents whose utilities imply random choosing, so respondents in the scale=0 class we count as random responders.

3.0 Evaluating Diagnostic Models

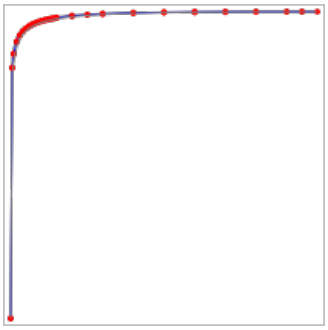
The classic tool for gauging the success of a diagnostic measure is the receiver operating characteristic (ROC) curve and an associated summary metric called AUC - the area under the curve. The ROC curve plots the true positive rate of a diagnostic measure against the false positive rate for different scores on a diagnostic measure. An ideal test will have a 100% true positive rate (in our case it will identify random responders) and a 0% false positive rate (in our case it will not mis-identify valid respondents as random responders), as represented by the upper left intersection formed by the blue lines on this ROC curve diagram:



The area under the blue lines is the entire square, so the blue curve's AUC is 1.00 (100%).

The orange line represents a perfectly worthless ROC curve – it has no diagnostic value because it represents the true positive rate always equal to the false positive rate (if the curve fell below the line, it would again have diagnostic value, because we could simply flip the scale so that a negative test score predicts a positive outcome). The orange line divides the area of the square exactly in half, so its AUC is 0.50.

In practice ROC curves fall between these extremes. For example, on this curve, while we can't get perfect prediction, it's possible to select a cutoff or threshold where we get a high true positive rate and a low false positive rate, because many of the points are close to the upper left corner that represents perfect prediction:



4.0 Two Artificial Data Studies

We base our analyses on two empirical studies, one CBC and one MaxDiff. The CBC experiment featured 6 attributes with 3-5 levels each (specifically it was a $5^3 \times 4^2 \times 3$ experiment) measured via choices in a dozen sets of triples. The MaxDiff had 36 items measured in 27 sets of quads. Each study had 2,400 human respondents. Unfortunately, we do not know which human respondents answered the survey questions diligently and which did so randomly.

We can, however, do this work with artificial respondents. First, we can program a set of 2,400 robotic respondents to make choices that conform to Random Utility Theory (RUM) using the

utilities we observed from our 2,400 human respondents. We know our 2,400 robotic respondents are valid, non-random respondents, because we program them to behave according to random utility theory, upon which the multinomial logit mathematical model and the applied choice modeling it supports both depend. So for each the CBC and the MaxDiff study we have 2,400 RUM-choosing robots whose utilities we know to be valid reflections of considered, theoretically appropriate choice behavior. Let's call these our Valid RUM respondents.

We can also program some number of artificial respondents to be random responders to the CBC questions. Let's call these our Latent Random responders, because we're going to mix them in with our Valid RUM responders and then see how well we can distinguish the two groups. We can then see how well each of our diagnostic methods in section 2.0 above performs with respect to identifying true positives (Latent Random responders) and false positives (RUM responders falsely accused of being random).

Finally, we can also program the extra random respondents needed for the RLH and LC methods above. For clarity we might call these our Diagnostic Randoms (because we generate them for diagnostic purposes only, as described in section 2.0 above).

As part of our experiment we created three versions of the CBC and three versions of the MaxDiff, to reflect different levels of sparseness: for the CBC treatments, we built experiments with 5, 10 and 15 CBC questions (i.e. wherein each respondent sees each level at least 3, 6 or 9 times, respectively) and for MaxDiff we built experiments with 9, 18 and 27 questions (i.e. wherein each respondents sees each item once, twice or three times, respectively). We also tested four different levels of Latent Random respondent incidence, 5%, 10%, 15%, and 25%.

5.0 Results

5.1 Preliminary findings

Early on in our analyses we learned some things about two of the diagnostic methods in section 2.0. First, in the LC method, we found instances where no class contained a majority of Diagnostic Randoms, giving us a 0% true positive rate. To prevent this from happening, we modified the method so that we classified as randoms (a) the Valid RUM or Latent Random respondents who were in classes where a majority of the respondents were Diagnostic Randoms, or (b) the Valid RUM or Latent Random respondents who were in classes where the largest proportion of respondents were Diagnostic Randoms

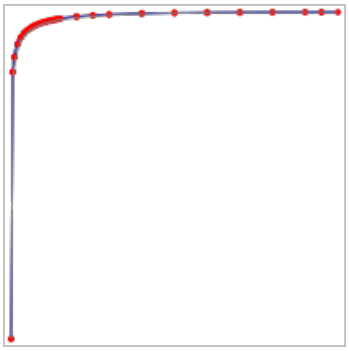
The other early learning concerned the RLH method. We found that the RLHs of Latent Random respondents are higher when their utilities are estimated together with the utilities of Valid RUM respondents than when they were estimated separately, as with the Diagnostic

Random responders. Thus, a cutoff value that identifies 95% of Diagnostic Random respondents only identifies about 80% of Latent Randoms (whose utilities are estimated together with the utilities of Valid RUM respondents).

We also learned early on that the incidence of Latent Random respondents didn't have a lot of impact on our ability to identify them, so we were able to simplify our analyses by collapsing our results across the four levels of Latent Random incidence.

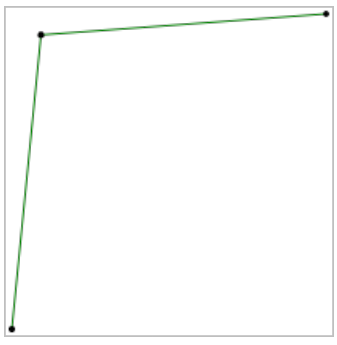
5.2 Substantive findings

For CBC, the richest design (15 questions), has an AUC of 0.988 for the RLH diagnostic:



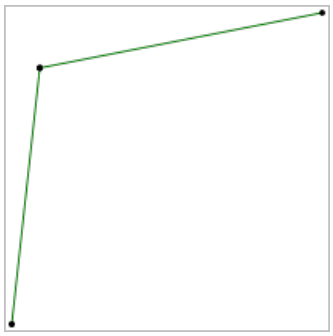
The RLH method is scalable and can be adjusted to allow for the potential different harms that come from including Latent Random respondents or from excluding Valid RUM respondents. Appendix 1 shows how many Latent Randoms and how many RUMS are misidentified at various cutoffs for our CBC experiment while Appendix 2 shows these results for our MaxDiff experiment. For example, at an RLH cutoff that eliminates 95% of Diagnostic Random responders (based on our sample of 1,000 Diagnostic Random respondents where the cutoff would be 0.482) we can eliminate 80% of the Latent Randoms at a cost of only 1.5% of the Valid RUM respondents.

For the LC method of identifying Latent Randoms, the AUC is a little worse, at 0.920:



Note that only one point on the curve can be used to discriminate RUM from random respondents, versus the continuum of points available with the RLH method, so we don't have the flexibility to look at different cutoff points.

AUC is worse still (0.866) for the SALC method and suffers from the same single point limitation as the LC method:



As you'd imagine, when we have sparser data, for example 10 CBC questions, the AUCs decrease (because we have fewer observations with which to do the discriminating): 0.957 for the RLH method, 0.871 for the LC method and 0.799 for SALC. By the time we get to the sparse condition where we ask only 5 CBC questions, we have AUCs of 0.856 for the RLH cutoff method, 0.716 for the LC method and 0.760 for the SALC method.

Results are similar for MaxDiff (we didn't run the SALC analysis for MaxDiff because we'd already found it to be inadequate for CBC):

- For the rich data condition (each item seen 3x per respondent), AUC is 0.991 for the RLH method and 0.744 for the LC method.
- For moderate sparseness (each item seen twice per respondent) AUCs are 0.977 for the RLH method and 0.668 for the LC method.
- And again, both fair much worse with sparse data (each item seen once per respondent): AUC of 0.809 for the RLH method and 0.561 for LC.

6.0 Summary/Recommendation

The RLH method dominates both of the methods based on latent class MNL. We find no reason to use the LC methods (except that the SALC method does not require that we create any artificial Diagnostic Randoms, which removes a step from the process, a simplification that harms our ability to distinguish valid from random respondents).

Using a 90% or 95% cutoff based on 1,000 Diagnostic Randoms, we can expect to be able to eliminate 75% or more of Random respondents while discarding in error potentially only 1-2% of valid RUM responders (though more of them as our design gets more sparse).

Appendix 1: CBC

Diagnostic Randoms Cutoff (%)	**Each level seen 9x/respondent** Randoms Identified (%)	RUMs Rejected (%)	**Each level seen 6x/respondent** Randoms Identified (%)	RUMs Rejected (%)	**Each level seen 3x/respondent** Randoms Identified (%)	RUMs Rejected (%)
75	60	<1	57	2	51	9
80	64	<1	62	2	55	10
85	69	1	68	3	63	13
90	76	1	74	4	67	15
95	80	2	80	6	73	18
99	89	3	91	13	87	35

This table shows true positives (Random respondents identified) and false negatives (RUM respondents rejected as random).

Appendix 2: MaxDiff

Diagnostic Randoms Cutoff (%)	**3 item views/respondent** Randoms Identified (%)	RUMs Rejected (%)	**2 item views/respondent** Randoms Identified (%)	RUMs Rejected (%)	**1 item view/respondent** Randoms Identified (%)	RUMs Rejected (%)
75	47	<1	46	<1	6	<1
80	54	<1	50	<1	9	<1
85	58	<1	55	1	12	<1
90	66	<1	62	2	19	2
95	76	1	72	3	33	3
99	90	2	88	6	53	7

This table shows true positives (Random respondents identified) and false negatives (RUM respondents rejected as random).

References

Hoogerbrugge, M and M. de Jong (2019) “Can we use RLH to assess respondent quality?” *Sawtooth Software Conference Proceedings, 105-112.*

Magidson, J. & J.K. Vermunt (2007) “Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference,” *Sawtooth Software Conference Proceedings*, 139-154.

Orme, B.K. (2019) “Consistency cutoffs to identify “bad” respondents in CBC, ACBC and MaxDiff,” https://www.linkedin.com/pulse/identifying-consistency-cutoffs-identify-bad-respondents-orme/?trackingId=g9ylG8GeasHXn79cZcC5JQ%3D%3D