Adaptive Survey Strategies: Optimizing Question/Answer Search

James H. Collins – GfK MRI

Introduction

A seemingly endemic quality of print audience readership research (and certainly other media and non-media survey research, etc.) is survey instruments (broadly conceived to include CAPI, CASI, personal interview, etc.) incorporating long sets of question/answer choices. Examples abound, but typically readership audience surveys involve multi-hundred item magazine/newspaper lists involving screens, recency and/or frequency and perhaps ancillary qualitative measures (e.g. How Obtained, A Favorite, Hardcopy/Digital, etc.).

Print audience measurement has not gotten to where it is with respect to survey length and its untoward consequences (broadly characterized as respondent burden) through thoughtlessness. Rather, it is where it is having traveled a long road beginning decades ago with the measurement of typically a few dozen large titles. Slowly but inevitably this road has gotten steeper and steeper with the introduction of title after title year after year as readership has fragmented and publishers have sought more focused audience markets. Readership research has tried to please all the people all the time (i.e. accommodate the various forces at play in this complex media market), but in that pursuit lurks the potential for death, or at least profound disability, by a thousand cuts.

Now of course, print audience research has evolved. In fact much of the record of WRRS/PDRF is the history of significant and sustained efforts (successful or not) to address what can be termed as measurement inflation? Perhaps flawed are some of these attempts, but willful ignorance and neglect are certainly not proximate causes.

Broadly, print researchers have adopted a number of different approaches to deal with this challenge, some directed toward re-engineering the user experience to allow for more efficient navigation through the survey (e.g. CAPI and CASI), others adopting one form or another of (more or less intelligent) versioning followed by some sort of post-collection imputation.

Choice modeling provides a fresh direction. Naïve conjoint analysis, wherein the number of potential choice options increases exponentially, is impractical in all but the most limited exercise. Hence, conjoint and choice-based techniques generally have evolved to incorporate adaptive strategies implemented online or within some other interactive and responsive computing environment. While particulars vary, generally these adaptive strategies use prior choice responses (from the individual respondent or a portion of the completed sample) to determine subsequent choice options thereby limiting the expanse of the questioning for any particular respondent. By design the full (or near full) option space is explored by the sample aggregate, but individual respondents are exposed to only a limited and (usually) the most relevant region of this full space.

The focus of the work at hand is the evaluation of a variety of different adaptive questioning strategies in the context of long media (magazine, cable network, internet) question/answer lists. As such, this work involves aspects of both of the strategies print audience research has utilized (as noted above) to render its tasks more tractable – versioning and technology to enhance the user experience. In the work at hand, the versioning is adaptive, i.e. guided by the respondents' responses, and involving technology to determine the appropriate reduced question sequencing.

The basic design of this evaluation is as follows:

1) Print audience readership from GfK MRI's Spring 2016 release of ~24,300 respondents will serve as the foundation for the simulations and analyses. This dataset is fully known and complete with respect to print readership (screens, reads and frequency) and demography, etc. and hence is appropriate for both simulation and evaluation purposes.
2) Employing this GfK MRI dataset a variety of different informed adaptive strategies will initially be evaluated. Basically, these strategies will involve splitting the GfK MRI dataset into two groups – Foundation and Simulation. The Foundation group will constitute prior information from which will be developed various association matrices to guide the simulation employing the second separate Simulation group. These matrices can be thought of as different strategies for selecting subsequent question/answer options based upon respondents' prior responses.
3) Using the Simulation group, for reduced question/answer sets, evaluate these various association matrices/strategies with reference to completeness of coverage of the known positive responses. In short, how efficient is the strategy with respect to selecting questions/items for which respondents have higher probabilities of positive responses?
4) The analysis will be further enhanced to include model-based approaches (regression, Bayesian Network, Association Rules) to item selection.

With respect to the variety of different basic transition matrices/adaptive strategies evaluated:

1) Random selection – used as a reference and analogous to naïve versioning.

2) Proportional selection – based on simple probability of positive response
3) Correlation – the next question/item is chosen based upon its (Yes=Positive or No=Negative) correlation with respondents' answers to prior question(s)/item(s).
4) Page Rank – as above with Correlation, but using a matrix/strategy derived from Google's basic search algorithm as developed by Brin and Page.
5) Regression – as above but using Linear Regression with independent variables determined by prior items and responses.
6) Bayesian Network – as above but using a trained (from the Foundation set) Bayesian graphical network.

In summary, the ambition is to identify dynamic questioning strategies both feasible for implementation in online surveying environments and efficient (and intelligent) with respect to coverage of discriminating measures.

Basic Adaptive Strategies

The basic adaptive strategies all follow the same design:

1) Assume a known and complete set of question item responses. In the case at hand readership of 174 magazines and national newspapers for ~24,300 respondents to the Spring 2016 GfK MRI Survey of the American Consumer.
2) Using the screen, read and frequency measures for each publication, compute the empirical probability of reading for each publication for each respondent.
3) Select at random from this probability dataset some number of respondents to constitute the Foundation dataset. From this Foundation dataset construct a number of different association/ transition matrices, wherein the probability of selection of the next item is based on the response to the current item. This transition matrix has Markov-like properties albeit with a 0 (zero) probability of returning to a prior condition – in short, the selection probabilities change as the process continues, i.e. are adaptive.

For the purposes of this evaluation a number of differently constructed association/transition matrices were constructed – these constitute the initial adaptive strategies and are relatively straightforward. Following are the various transition matrices and how they are constituted:

1) Random – Each publication has the same probability of selection. This comprises the benchmark against which all other adaptive techniques are evaluated.
2) Audience Size – Each publication has a probability of selection proportional to the size of its audience.
3) Audience Size Squared – Each publication has a probability of selection proportional to the squared size of its audience.
4) Correlation – The probability of selection is based on the correlation matrix for the publications, i.e. the probability of selection of the next publication is related to its correlation to the prior publication.
5) Correlation Plus – Same as Correlation above, but if the respondent in the simulation has not read the prior publication the selection of the subsequent publication is at random.
6) Positive/Negative Correlation – Similar to Correlation above, but in the case wherein the respondent in the simulation has not read the prior publication the probability is based on the correlation of not reading the prior with reading the subsequent.
7) Page Rank – The probability of selection is based on the Page Rank association measure as developed by Sergey Brin and Larry Page ["The Anatomy of a Large-Scale Hypertext Web Search Engine"] and employed by Google (in heavily enhanced and refined fashion algorithms) in its search engine.

A series of simulations were run with varying sizes of a) the Foundation respondent base upon which the association/ transition matrices were developed and b) number of publications to be evaluated. With respect to the first, the primary question was, "How large does the base of respondents for whom there is complete information (i.e. all publication asked) need to be such as to develop reliable association/transition matrices?" With respect to the second, the primary question was, "How many publications does each respondent need to be exposed to so as to capture a significant share of their reading?"

The structure of the simulations employing these various transition matrices is relatively straightforward particularly given the know outcomes for the readership items in the Simulation group of GfK MRI respondents. The simulation proceeds as such:

1) For each GfK MRI respondent in the Simulation group a publication is chosen at random, using either a uniform or some other probability distribution appropriate to the transition scheme being simulated.
2) For each respondent in turn, using their known response to this first publication the subsequent publication is chosen using the probability distribution embodied in the association/transition matrix.
3) Step #2 proceeds for each respondent until the maximum number of publications to be asked has been reached. It is in this fashion that the process is adaptive; the probability of particular successive print items being asked is dependent upon the response to the proceeding item(s).
4) Note, unlike a standard Markov process, once a particular state has been reached it is assigned a 0 (zero) probability of being revisited, i.e. once a respondent is asked about a particular publication that publication is not asked again.

The efficacy of these various adaptive schemes was evaluated by a variety of different metrics, but generally in comparison with the Random scheme noted above. In brief the evaluation criteria amounts to, "How much better than random is the particular adaptive technique with respect to selecting publications about which to ask the respondents that in fact they have screened/read?"

Tables #1 and #2 respectively report the results for seven basic adaptive techniques for which 50 (Table #1) and 100 (Table #2) of the total of 174 publications were asked. In these examples the Foundation dataset was comprised of only 1% of the total ~24,300 GfK MRI Spring 2016 respondents and the Simulation dataset of the remaining 24055 respondents. This is important insofar as the transition matrices need complete information (i.e. all publications), but as is evident from the results it is not critical that this foundation dataset be large.

Table #1 – 50 Publications Asked with Foundation of 240 Respondents (1%)

Simulation	Mutations	Percent of Total Screens	Percent of Total Reads	Quality Score
Random	0	29	29	1.00
Random	9	29	29	1.01
Audience Size	0	42	45	1.46
Audience Size	9	42	45	1.45
Audience Size Sqr	0	48	53	1.67
Audience Size Sqr	9	47	52	1.65
Correlation	0	42	46	1.47
Correlation	9	42	45	1.46
Correlation Plus	0	42	45	1.47
Correlation Plus	9	42	45	1.46
Pos/Neg Correlation	0	48	53	1.68
Pos/Neg Correlation	9	48	52	1.66
Page Rank	0	44	47	1.54
Page Rank	9	44	47	1.53

Table #2 – 100 Publications Asked with Foundation of 240 Respondents (1%)

Simulation	Mutations	Percent of Total Screens	Percent of Total Reads	Quality Score
Random	0	58	58	1.00
Random	9	59	59	1.03
Audience Size	0	72	75	1.25
Audience Size	9	74	77	1.29
Audience Size Sqr	0	75	78	1.31
Audience Size Sqr	9	80	83	1.39
Correlation	0	72	75	1.26
Correlation	9	75	77	1.30
Correlation Plus	0	72	75	1.26
Correlation Plus	9	74	77	1.29
Pos/Neg Correlation	0	76	79	1.32
Pos/Neg Correlation	9	80	84	1.40
Page Rank	0	74	76	1.29
Page Rank	9	77	79	1.34

Beginning with Table #1 (50 publications asked) as expected the Random process captured only 29% or the Screens and Reads – 50 is 29% of the total 174 – and it is against this level that the genuinely adaptive schemes are evaluated. All of the non-random schemes surpass random by at least 45%+ with the Positive/Negative Correlation scheme performing best.

Note also that each strategy was run twice yielding slightly different results. An artifact of the non-random schemes is that they will tend to select the high probability (i.e. large audience) publications. This makes sense and is largely desirable as these techniques are designed to ask about publications wherein respondents have high(er) probabilities of responding positively and large publications de facto meet this quality. With that said, it is important that coverage extend to smaller publications, de facto with smaller probabilities of exposure and hence smaller probabilities of selection even using adaptive means. To address this matter, in the simulations involving non-zero (9 in this case) mutations, ~5% of the publications had their probabilities adjusted at random as the next publication was selected for each respondent. The design of this mutation strategy was to randomly select publications for mutation based on their inverse probabilities of selection and replace them with probabilities for publications based on their probability of selection. In short, the tendency was to raise the probabilities of low(er) probability/smaller publications and lower the probabilities of high(er) probability/larger publications. By way of reference this technique is reminiscent of strategies employed in Genetic Algorithms to insure optimization search space is reasonably covered – reasoning analogous to the case at hand. We can see in particular that this mutation strategy is beneficial in the Correlation Plus, Positive/ Negative Correlation and Page Rank adaptive strategies (the most aggressive strategies) in the case of 100 publications being asked and does essentially no harm in the 50 publications case.

Finally with respect to these initial simulations of the adaptive schemes the Foundation was raised from 1% of the GfK MRI respondent base (~240 respondents) to 50% of the base (12054 respondents). Table #3 shows that this significantly larger poll of complete respondents makes essentially no difference to the effectiveness of the adaptive schemes. This is encouraging insofar as the central purpose of the adaptive strategy is to not have to burden respondents with large numbers of items, and as the Foundation set utilized in these initial simulations requires such is best minimized.

Table #3 – 50 Publications Asked with Foundation of 12054 Respondents (50%)

Simulation	Mutations	Percent of Total Screens	Percent of Total Reads	Quality Score
Random	0	29	29	1.00
Random	9	29	29	1.01
Audience Size	0	42	45	1.45
Audience Size	9	42	45	1.44
Audience Size Sqr	0	48	53	1.67
Audience Size Sqr	9	47	52	1.65
Correlation	0	42	46	1.46
Correlation	9	42	45	1.45
Correlation Plus	0	42	45	1.45
Correlation Plus	9	42	45	1.45
Pos/Neg Correlation	0	48	53	1.69
Pos/Neg Correlation	9	48	52	1.67
Page Rank	0	44	47	1.53
Page Rank	9	44	47	1.52

Model-Based Adaptive Strategies

The adaptive strategies heretofore employed each perform fairly well when compared to random selection and have the nice property that they are efficient and straightforward to implement in an interactive (e.g. online) survey environment. Moreover, when implemented with some measure of mutation they are relatively expansive in their search (i.e. the space of low incidence print vehicles is reasonably considered) and maintain their performance levels. With that said, they do not utilize the histories of selection as much as possible. What is meant here is that the entire chain of a respondent's answers is potentially informative with respect to subsequent questioning, but the schemes considered thus far largely base the subsequent choice only the immediately prior choice (or the subsequent choices are independent, e.g. Audience Size). Now this is a somewhat exaggerated charge insofar as all prior choices are given zero probability of subsequent inclusion thereby altering the probabilities of the remaining choices. Nevertheless the force of the point persists – the informative history of all prior choices is limited.

Hence, consideration of model-based alternatives attempting to more fully employ this informative history is warranted.

A simple mechanism, at least in design, is to employ regression modeling wherein a respondent's responses to prior selections (i.e. their history) constitute the independent variables in the regression and each available selection the dependent variable in a series of models. The scheme is constructed as follows:

1) For each respondent in succession, the first selection is based on some method requiring no prior selection (e.g. Audience Size).
2) With the first selection made a series of regression models are run with the first selection as the independent variable and each remaining possible selection the dependent variable in the series of models.
3) The respondent is then scored for each model in the series and the propensities from this series of regressions constitute the probability distribution from which the next selection is made.
4) The set of independent variables increases by one to include the last selection and the series of regressions decreases by one as the last selection is no longer available as a dependent variable.

5) Again the respondent is scored for each of regression model in the series and the next selection made as noted above.

Because of the computational intensity of this approach (i.e., thousands of separate regressions for each respondent) the simulation was limited to only 1% of the Simulation set (240 respondents). To better establish a reference this regression scheme was compared to both the Random and the Page Rank schemes each run on the 1% Simulation sample. Table #4 and #5 shows the results for 50 and 100 publications respectively.

Table #4 – 50 Publications Asked with Foundation of 240 Respondents (1%) (Simulation Set of 240)

Simulation	Mutations	Percent of Total Screens	Percent of Total Reads	Quality Score
Random	0	27	27	1.00
Random	9	27	28	1.01
Regression	0	34	34	1.26
Regression	9	34	35	1.25
Page Rank	0	42	42	1.56
Page Rank	9	42	42	1.50

Table #5 – 100 Publications Asked with Foundation of 240 Respondents (1%) (Simulation Set of 240)

Simulation	Mutations	Percent of Total Screens	Percent of Total Reads	Quality Score
Random	0	54	54	1.00
Random	9	54	53	0.99
Regression	0	62	60	1.11
Regression	9	64	63	1.19
Page Rank	0	69	69	1.28
Page Rank	9	73	72	1.36

In short, the regression strategy, while performing better than random substantially underperforms the relatively successful Page Rank method.

But why? Much has to do with the size of the Foundation set – 240 respondents. While this sample size supports quite well the estimation of the relationships between pairs of publications in all likelihood it is insufficient to support models having a more multi- variant character.

Table #6 below shows results for a simulation identical to that reported in Table #5 but with the Foundation set of ~12,000 respondents. The results are considerably improved for both the Regression and Page Rank strategies.

Table #6 – 50 Publications Asked with Foundation of ~12,100 Respondents (~50%) (Simulation Set of 240)

Simulation	Mutations	Percent of Total Screens	Percent of Total Reads	Quality Score
Random	0	27	27	1.00
Random	9	27	28	1.01
Regression	0	47	48	1.78
Regression	9	45	48	1.71
Page Rank	0	44	47	1.74
Page Rank	9	44	47	1.68

Bayesian Network

Finally, a Bayesian network was evaluated albeit in a more modest fashion than the techniques previously considered.

As background, Bayesian networks are:

> "...[A] class of *graphical models* that allow a concise representation of the probabilistic dependencies between a given set of random variables...as a directed acyclic graph...Each node...corresponds to a random variable X_i." ["Bayesian Networks in R", Nagarajan, Scutari and Lebre, Page 13]

In the context of the problem at hand, a Bayesian network manifests the conditional probability structure of the magazine audiences. In the particular instance evaluated the Bayesian network was built from a discretized matrix of reading probabilities for the 174 print vehicles – (None = 0, Low <= 0.2, High > 0.2) – using the Hill-Climbing algorithm available in the R bnlearn package [Marco Scutari (2010). Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35(3), 1-22.].

As note previously the evaluation of the Bayesian network was more limited than the various other technique considered, this chiefly related to performance of the simulation. To elaborate, as noted above, the magazine reading probabilities were discretized into three categories (None, Low and High). For the purposes of actually constructing the Bayesian network, the continuous or discretized data is, practically considered, of comparable and feasible computational burden. However, it is with respect to "querying" the network (i.e. determining the probability of selection) that performance emerges as a major factor (perhaps better termed, "rears its ugly head"). The simulation as constructed (and to be detailed in subsequent paragraphs) took approximately 48 hours (tedious but nevertheless feasible) using the discretized data whereas the Gaussian/continuous network simulation was of undetermined but wholly untenable duration. Note, as a practical matter this is of only modest consequence, for in a practical application of an adaptive survey the process itself occurs over time (e.g. survey administered within relatively long survey period) and even the administration of the survey to a single respondent occurs at a leisurely pace relative to the speed of computation. In short, while for the purposes of the simulation computational constraints of employing Gaussian/ continuous Bayesian networks emerge prominently, in practical administration they are a very modest factor.

The simulation involved constructing a Bayesian network of the 174 print vehicles with a foundation of 6140 respondents. To render the simulation tractable the Simulation dataset included only 545 respondents, no mutation was employed and 50 publications were asked. In comparison with the Random print vehicle selection strategy the Bayesian network performed about 42.5% better with respect to reads. However, by way of comparison the Page Rank method performed 68.75% better.

Conclusions

While not exhaustive, a variety of different adaptive questioning strategies have been considered and in comparison with the Random method all have demonstrated substantial improvement in identifying those items for which respondents' are individually more likely to answer in the affirmative. What is promising is that each of these strategies is fairly simple to implement and computationally efficient (with the possible exception of the Bayesian network, which is potentially also feasible.)

With that said, the work at hand is perhaps best understood as an invitation to print media and other researchers to explore survey research designs and implementations which turn away from static, monolithic, procrustean methods and towards more responsive ones.

Author

James Collins is Executive Vice President, Research at GfK MRI where he focuses on data integration. A frequent contributor to PDRF and other conferences, Collins' primary research foci beyond data integration are optimization and systems development.