Responsive design in ONS Telephone Operations

Steve Woodland, Head of Survey Operations, Office for National Statistics and **Jen Farnall**, Senior Researcher, Office for National Statistics.

1. Introduction

Responsive design is "an adaptive data-collection approach that uses information available, both before and during data collection, to adjust the collection strategy for the remaining cases." [1] In most surveys all sample units receive the same treatment and the same design features apply to all cases. Responsive designs assume that different sample units may receive different treatments. These treatments are defined before the survey starts, but may also be updated via data that are observed during data collection.

The motivation to implement a responsive design came from the 2014 National Statistics Quality Review (NSQR) of the UK Labour Force Survey (LFS)[2]. The review recommended: "Explore the potential for adaptive fieldwork to focus resources more efficiently – in particular to: a) identify priority areas in advance of fieldwork; and b) explore opportunities for adaptive design during fieldwork." This paper focuses on the second of these recommendations where the aim was to determine whether data collection can achieve the same response rate more efficiently whilst also ensuring non-response bias is not negatively affected.

The LFS is the UK's largest household survey. Interviewing takes places over five consecutive quarters (waves) with the main mode of interviewing being face to face (CAPI) at wave 1 and telephone (CATI) for subsequent waves (waves 2-5). Capability to implement a responsive design given the current IT architecture was greatest for the wave 2-5 cases. The functionality of the Blaise CATI call scheduler would allow the management of cases according to pre-determined rules.

Currently no demographic information about a household is used in prioritising cases. LFS cases are scheduled so that every case starts with an equal chance of being called. Prioritisation rules, based on a case's call history in that wave so far, are then used to determine priorities for subsequent calls. A program is run every five minutes to re-prioritise remaining cases.

The field period of the LFS is just over two weeks and each week there will be three weeks of data collection proceeding simultaneously for different cases. At any given time there will be:

- cases that haven't been tried since last wave (week 1 cases);
- cases that have already been tried for one week (week 2 cases); and
- cases that have already been tried for two weeks (week 3 cases).

To investigate responsive design, only week 2 cases were included in the scope for analysis. This means that only cases not yet contacted by week 2 of the field period will be subject to the responsive design intervention. Any cases that have already completed an interview, already have an appointment scheduled for week 2, or have already refused will not be in scope.

2. Methodology

The approach we have adopted is to calculate a response propensity for each case and then prioritise cases based on their likelihood of responding, with the aim of achieving the same response rate with less interviewer resource. However, although the project does not aim to reduce non-response bias it is still important to ensure that non-response bias does not increase because of any design implemented. The

[1] **Responsive Collection Design Framework for Multi-Mode Surveys**, Francois Laflamme, Statistics Canada, April 2013

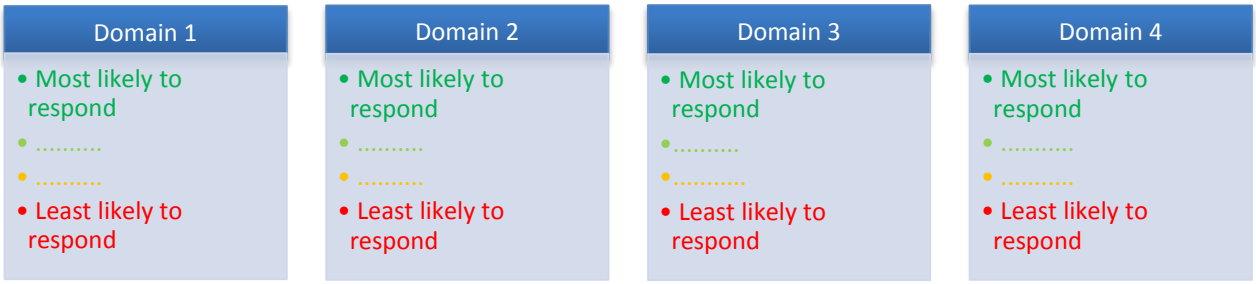
[2] http://www.ons.gov.uk/ons/guide-method/method-quality/quality/quality-reviews/list-of-current-national-statistics-quality-reviews/nsqr-series--2--report-no--1/report---review-of-the-labour-force-survey.pdf

design therefore ensures that groups which have the potential for under-representation - either because they are under-represented at wave 1 of the LFS, or because they are more likely to drop out of the survey between waves - are not ignored. These groups need to be prioritised equally with cases that are over-represented or less likely to attrite. Under-represented cases are likely to have a low probability of responding and would therefore be given a low priority within the call scheduler if the design were based purely on propensity modelling. This could then result in greater under-representation and cause further non-response bias in the sample than is already present.

To overcome this, the design allocates sample units to pre-defined mutually exclusive domains before any prioritisation is done. These domains are based on response rates of different groups, where the groups are put together based on household characteristics, and take into account representativeness at wave 1, as well as representativeness after attrition. The only information available on the non-responders at wave 1 is what can be found from the Census Non-Response Link Study (CNRLS)[3].

The domains are therefore based not only on response rates for different groups at waves 2+ (i.e. taking account of those most likely to attrite) but also take into account factors which the CNRLS identified as a concern at wave 1 (i.e. taking account of household types which are less likely to take part at all). Additionally, domains are fixed groups that households can be allocated to based on their household characteristics, rather than using any information gathered from survey questions that is likely to change from one wave to the next (employment status for example). All domains are weighted equally and propensities to respond are used to prioritise within the domains. So for example, a case which is the most likely to respond out of all cases in domain 1 is weighted equally against a case which is the most likely to respond out of all cases in domain 2, even if they have very different absolute likelihoods of responding. By following this approach, non-response bias should not increase.

A model has been developed based on information from the previous wave(s) and from the first week of data collection in the current wave to calculate the propensity of each sample unit within each domain responding in the second week of data collection. Interviewer resources will be focused on the units most likely to respond within each domain to try to maximise response within each domain. That is, the Blaise CATI scheduler will prioritise those cases most likely to respond within each domain, so these cases will be presented to interviewers first. This means that in week 2 of the field period, resources will be focused on the units most likely to end in success, in domains where representation or the overall likelihood of responding is low, as well as in groups which are well represented or where the overall likelihood of responding is high. The diagram below illustrates this approach. How many cases will be worked on in week 2 will depend on resources. It might be that those least likely to respond never get tried, or may be given a different treatment altogether, for example being reissued to field or being sent a letter.



To establish the domains and develop a propensity model, call record data for Wave 2 LFS cases in 2012 and 2013 were linked to their survey data and call history information from Wave 1. This gave information on household characteristics and previous wave call history for the responders and non-responders in Wave 2[4].

[3] The CNRLS linked 2011 Census data to wave 1 LFS data to find out information about the non-responders. http://www.ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/articles-and-reports/non-response-weights-for-the-uk-labour-force-survey.pdf

[4] Approximately 53,000 cases in scope.

2.1. Establishing domains

The characteristics that were looked at for assigning cases to domains were:

- Output Area Classification (OAC)[5]
- Ethnicity of Household Reference Person (HRP)
- Household type
- Tenure

These were all factors which the CNRLS identified as associated with LFS non-response bias. Response rates in wave 2 (of those who took part in wave 1) were calculated for all of these factors to ensure that groups among which we might expect to lose disproportionate numbers of respondents are not excluded. The aim was to try and get the smallest variance of response rates within each domain.

For each of the variables looked at, there is a clear group which had a lower response rate. These are:

- OAC – Supergroups 1 (Blue Collar Communities), 5 (Constrained by Circumstances) and 7 (Multicultural)
- Households where the Household Reference Person (HRP) is of an ethnic minority
- Households which are Lone Parents
- Households which are being rented

Testing whether the means and variances are equal has led to six distinct domains being established where the means of response rates across domains are significantly different and the variances within domains are small. The six domains are shown in the table to the right:

	HRP ethnicity	Household type	Tenure	OAC Supergroup
Domain 1	Ethnic minority	All	All	All
Domain 2	White	Lone parent	Renting	All
Domain 3	White	Lone parent	Not renting	All
Domain 4	White	Not lone parent	Renting	All
Domain 5	White	Not lone parent	Not renting	1, 5 or 7
Domain 6	White	Not lone parent	Not renting	2, 3, 4 or 6

The error chart to the right shows the means and variances of the six domains.

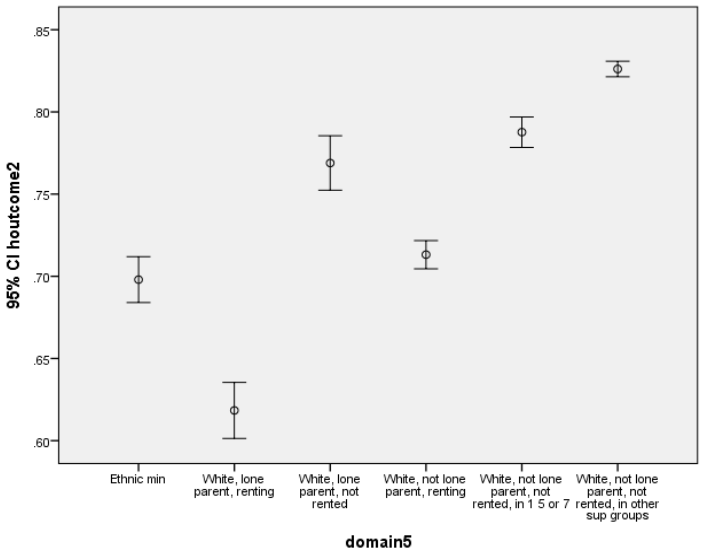


Chart 1: Error chart for chosen domains

Note: houtcome2 represents response where 1 is a success and 0 is not

[5] The Output Area Classification (OAC) system is a geo-demographic classification system which has been constructed by creating a hierarchy of clusters, which together typify the characteristics of a given area. The Output Areas of the 2001 Census were classified into 52 clusters, also called **subgroups**, which aggregate to 21 **groups**, which then further aggregate to 7 **super-groups**. To classify over 200,000 Output Areas into these groups and clusters, 41 attributes were used. For example, the "City Living" Supergroup has a combination of characteristics that you might expect from city areas, such as larger proportions of flats and private renters. Conversely, the "Countryside" Supergroup has a combination of characteristics like those you might expect to find in more rural areas, such as large proportions of detached housing, households with two or more cars etc. It is the combination of these characteristics that generates the distinct differences between the Supergroups.

2.2. Developing a propensity model

To develop a propensity model, the linked historical data was used to select the cases that would be in scope (i.e. those that had not been contacted by the end of week 1[6]) with what actually happened to them in weeks 2 and 3. The information from wave 1 (call history and household characteristics) and the call history from week 1 of wave 2 was used to develop the model. This information could be used to infer how likely a case was to respond in week 2 or week 3.

For the purposes of this analysis a logistic regression model was used to determine which variables are significant in predicting the final household outcome when it is known what has happened so far in week 1.

A forward stepwise method was used, so that at each step, the variable with the largest significant score (where significance is where the significance value $p<0.05$) is added to the model. Several models were investigated, and for each model, variables were added, removed or grouped depending on what the previous model had shown. Where predictor variables were categorical, a variable value was chosen as the indicator for other variable values to be compared against. If one value is significant in predicting response compared to another, then it will show as significant in the model. The output also shows if the overall variable is significant as well as the individual values. Four different tests of good fit were used to assess the suitability of each model and the predicted probabilities for each were plotted to look for any cases that were not being fitted well.

It was also important to consider how a model could be used operationally within the call scheduler. Only demographic variables that can be rotated forward from the previous wave can be used in a model.

The model that was chosen to be the best fit included the variables (in order of strength in the model):

- Total calls this wave so far (in bands)
- Tenure (Rented or not rented)
- Age group of HRP (in bands)
- Number of days the household has been tried this wave
- How many calls it took to contact the household last wave (in bands)
- Number of time slots the household has been tried this wave
- Household type (Lone parent or not lone parent)

The section below describes how this design will practically be implemented.

3. Implementation

To implement this design into the call scheduling system the processes will be as follows:

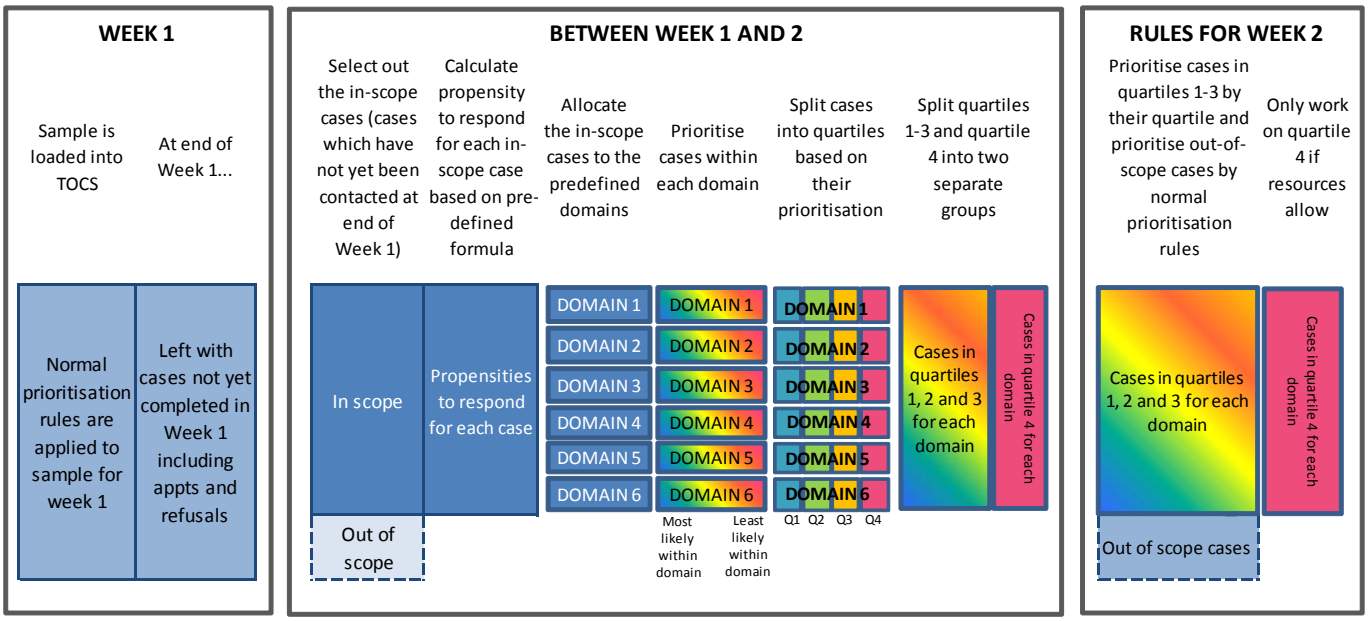
- Allocate cases into groups based on whether they are in scope or not.
- For the in-scope cases calculate a propensity to respond based on the model described in the previous section. The model will not be run every week with new cases, instead, a formula to calculate the propensity has been produced based on the results of the model above. This can then be programmed into the survey interviewing software for each case.
- Allocate in-scope cases to the pre-defined domains.
- Rank cases within each domain to quartiles, where quartile 1 are the 25% of cases most likely to respond within a domain and quartile 4 are the 25% of cases least likely to respond within a domain.
- Filter out the lowest quartile of each domain into another group, which will only be worked on if resources allow.

[6] There were around 7,500 cases that were eligible for this analysis

- Prioritise the remainder of the in-scope cases by their quartile rank. The out-of-scope cases will be prioritised just as they are currently, based on their call history in week 1.

The managers in Telephone Operations will have to allocate resources to the different groups (in-scope and out-of-scope), and the idea is that the same amount of resource should be given to the week 2 cases as what happens currently.

The diagram below shows the implementation process:



4. Evaluation

The responsive design will be piloted in TO in the October-December 2015 quarter, where the implementation processes will be tested. The design will then be implemented properly from January 2016.

The intervention will be considered a success if either

(i) the final response rate for in-scope cases is higher than would have been expected for in-scope cases before the intervention or

(ii) the response rate for in-scope cases is equal to the response rate that would have been expected for in-scope cases before the intervention and is achieved earlier on in the second week of data collection.

To evaluate this, the current profile of response is being recorded for TO LFS main cases for Waves 2-5. The profile records how quickly and how consistently response builds up with the current data collection design. This will allow comparison with the new profile once the responsive design is implemented.

Additionally, the design aims not to negatively affect the non-response bias in the LFS. To evaluate whether this has happened, the characteristics of respondents in Waves 2-5 main LFS will be looked at for the quarter where the design is implemented (JM16) compared to the same quarter last year (JM15). The proportion of households in each of the pre-defined domains will be used as a way of assessing whether the profile of respondents has changed because of the intervention.

Appendix

The chosen model is shown in the tables below.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 7[g]	Total calls this wave	-.559	.029	367.873	1	.000	.572
	No of days tried this wave	.124	.025	23.959	1	.000	1.131
	No of timeslots tried this wave	.184	.039	21.891	1	.000	1.202
	Tenure	.604	.061	98.568	1	.000	1.830
	Age group	.171	.020	74.478	1	.000	1.186
	No of calls to contact last wave	-.090	.019	22.116	1	.000	.914
	Household type	.162	.079	4.231	1	.040	1.175
	Constant	-1.396	.185	57.166	1	.000	.248

a. Variable(s) entered on step 1: Total calls this wave

b. Variable(s) entered on step 2: Tenure

c. Variable(s) entered on step 3: Age group

d. Variable(s) entered on step 4: No of days tried this wave.

e. Variable(s) entered on step 5: No of calls to contact last wave.

f. Variable(s) entered on step 6: No of timeslots tried this wave

g. Variable(s) entered on step 7: Household type.

Total calls is grouped into 1: 3 calls or less so far 2: 4 – 6 calls so far 3: 7 to 9 calls so far 4: 10-12 calls so far 5: 13-15 calls so far 6: 16-19 calls so far 7: 20 or more calls so far	**Age group** is: 1: Under 25 2: 25-34 3: 35-44 4: 45-54 5: 55-64 6: 65-74 7: 75+	**Calls to contact** last wave is grouped into 1: 1 call to contact 2: 2 calls to contact 3: 3 calls to contact 4: 4-6 calls to contact 5: 7-10 calls to contact 6: 11-14 calls to contact 7: 15 or more calls to contact	**Household** is 1: Lone parent, 2: Other
Tenure is 1: Renting, 2: Other	**Days** is a number from 1 to 7	**Times** is a number from 1 to 4	