Optimising allocation of households to web

Understanding Society IP5

|  |
| --- |
|  |
| Authors: Richard Boreham  Date: Sep 2013 |

# Understanding Society IP5

Within the Understanding Society project, there is an annual IP (Innovation Panel) for methodological testing. IP5 was the first wave to collect data in both web and FTF (face-to-face) modes.

There was a split sample experiment in IP5, where 33% of the sample was issued directly to FTF, and the remainder was initially issued to web and then transferred to FTF. Understanding Society is a longitudinal household survey with questions asked of the whole household and then separate instruments for each adult aged 16+[[1]](#footnote-1). Households were transferred to FTF unless all individuals in the household had completed their questionnaires.

Half of the sample that was initially issued to web was offered an additional incentive of £5 per adult if the whole household completed their questionnaires on the web before a certain date. The analysis in this paper just compares the web sample with the additional incentive to the sample issued directly to FTF to avoid confounding effects.

There are two main measures of response that we need to consider. The key target is the individual level re-interview rate for people who were interviewed at the previous wave. The other key figure is the full response rate for households completing via the web as this means that no interviewer costs would be incurred, and hence maximises cost savings.

34% of individuals completed their questionnaires online, and 25% of households were fully productive on the web. The overall individual response among the web-issued sample (once non-productive individuals had been transferred to FTF) was 70%, compared with 75% among the FTF issued sample.

The long term target for Understanding Society is to be able to issue a sizeable proportion of households initially to web and to get a 50% full household response rate (double the rate achieved on IP5), without damaging the overall individual level response.

# Theoretical approach to modelling response

We want to try to identify households which are more likely to have a full household response. The standard way that this is done is to create a household level logistic regression model to predict full household response. The problem with that approach is that you have to create household level composite variables from the answers that individuals have given.

This approach means that you lose some of the detail as you create your household level variable. Suppose that you think age might predict response, you have to create a household level variable that summarises the age of all the adults in the household. The easiest variable to create would be average age, but that would classify a household of two adults aged 40 with children the same as a 20 year old living with their 60 year old grandparent which seems intuitively wrong. The other method would be to group households into classification types, but that is also difficult because of the number of different household composition of ages that you come across.

The alternative approach that we have taken is to develop a model of individual response, and then to predict household response from the outputs of this model. This allows us to keep the granularity that we would otherwise lose by creating household level variables.

We randomly divided the web-issued sample in half[[2]](#footnote-2), and created a model on one half of the sample. Then we applied it to the other half to select which households we would choose to issue to web, and it is then possible to measure the actual response rates among this selected sample. We were also able to apply the same model to the sample that was issued directly to FTF to measure estimated FTF response rates[[3]](#footnote-3).

When measuring full household response, we have assumed that full response is interviewing all the people that gave an individual interview at the previous wave[[4]](#footnote-4).

# Model to predict response

The logistic regression model to predict individual response at IP5 among productive individuals from IP4 had the following significant and non-significant predictors.

|  |  |
| --- | --- |
| **Significant**   * Marital status * Tenure * Has valid email address * Frequent web user * Household outcome at IP4 * Number of productive individuals in household at IP4 | **Non-Significant**   * Age * Sex * Employment status * Likely to move in next year (IP4) * General health * Whether productive at IP1, IP2 or IP3 * Number of waves at which productive * Household size |

However we are not interested in which factors are significant per se, we want to apply this model to predict individuals’ propensity to respond. From that we can work out the probability of the whole household responding by multiplying the individual probabilities together[[5]](#footnote-5). Although the maths to estimate the probability of households taking part has its flaws, nevertheless we are interested in whether we can distinguish between households that are more or less likely to fully co-operate, rather than have an exact accurate prediction of the probability that they would be fully co-operating.

Once we have the estimated probabilities of fully co-operation for every household, we can rank order households in descending probability of response. Then we can look at the top 5%, top 10%, top 15% of households on this list and work out what their actual IP5 response rates were.

# Response rates by percentile of sample selected

|  |
| --- |
| 1. Web Individual Response, by Percentage of Households Issued to Web |
|  |

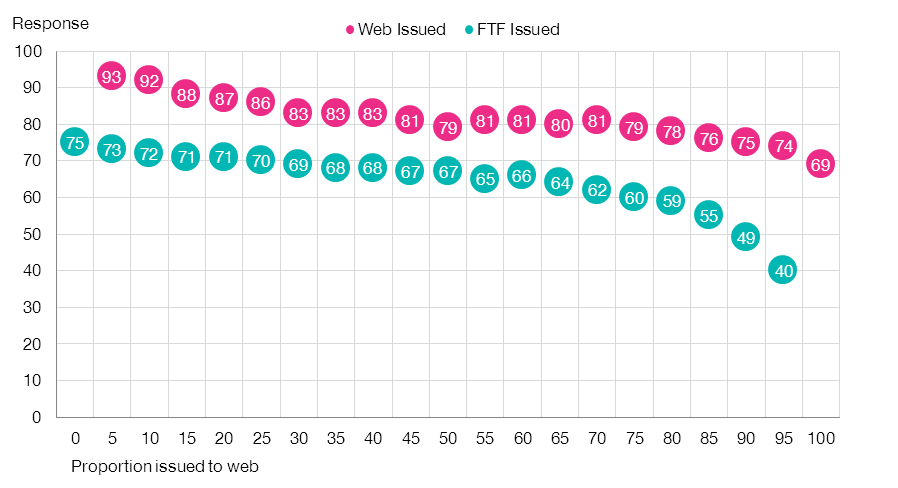
The chart above shows that we can select households which contain individuals that are more likely to respond. When we select the top 5% of households identified by our model, we get an individual response rate of 93%. The individual response rate among the top 10% of households is 81%, and among the top 15% is 78%. When the whole sample is issued to web the individual response rate is 35%.

Individuals who did not respond to web were transferred to FTF, and the following chart shows the overall response rates by the percentage of households issued to web.

|  |
| --- |
| 2. Web & Total Individual Response, by Percentage of Households Issued to Web |
|  |

Here we can see that the overall decline in total individual response rate in not as steep as the decline in web response by the proportion of households issued to web.

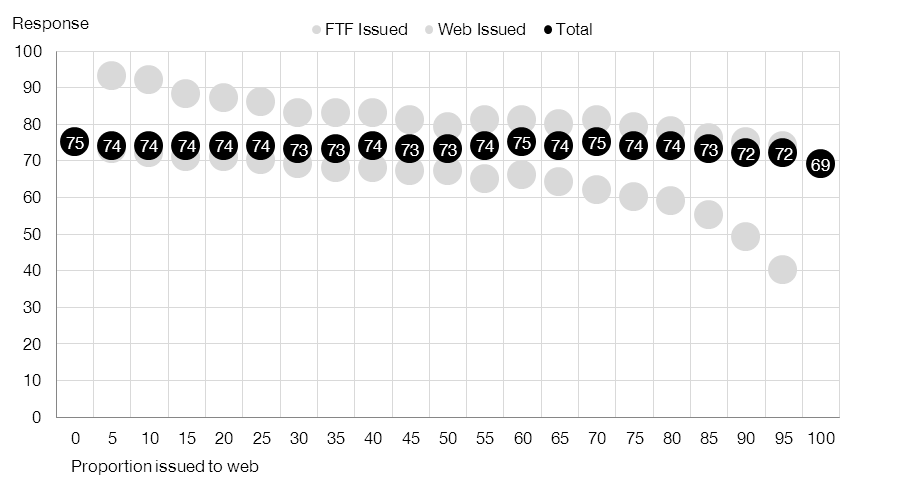
|  |
| --- |
| 3. Total Individual Response for Web Issued v FTF Issued, by Percentage of Households Issued to Web |



At first sight this chart looks counter intuitive, as overall individual response rates decline among both the web-issued and FTF-issued samples as the proportion of households allocated to web increases. When all households are issued to FTF, the individual response rate is 75%. When we cherry pick the top 5% of households to issue to web, this leaves the “worst” 95% to issue to FTF, and hence the individual response among FTF issued households is slightly lower than when all households are issued to FTF.

The figure that we are most interested in is the overall individual response rate combining the web-issued and FTF-issued samples.

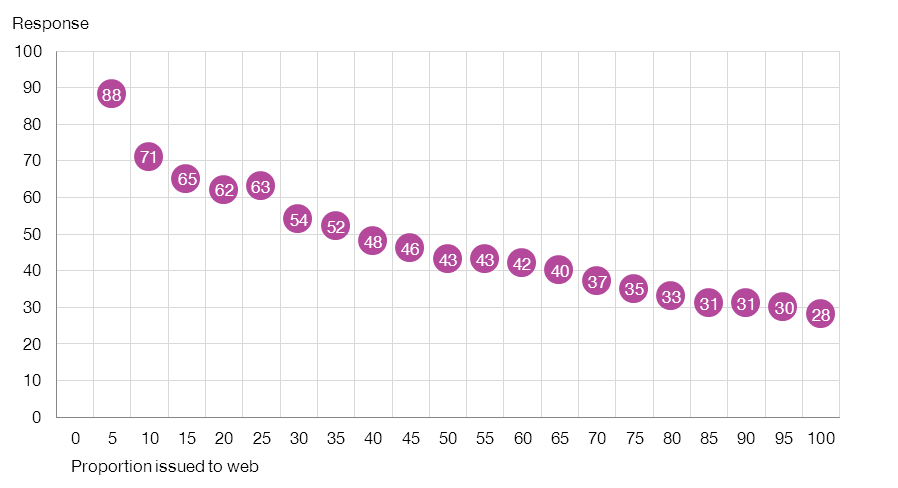
|  |
| --- |
| 4. Total Individual Response, by Percentage of Households Issued to Web |



Here we can see that the total individual response for the combined web and FTF issued samples is 75% when no households are issued to web and is 69% when all households are issued to web, and declines very slowly in between. This seems to suggest that until you issue over ¾ of households to web, that the proportion that you decide to issue should not impact on the overall individual response rate of the whole sample.

So it appears that we can choose households to issue to web where individuals are more likely to respond and not impact on the overall individual response rate. We will only make cost savings if we manage to have fully productive households in web as this means we can avoid sending an interviewer to visit that household. The final following chart looks at full household response.

|  |
| --- |
| 5. Web Full Household Response, by Percentage of Households Issued to Web |



This suggests that we would be able to select households which were more likely to be fully productive, and that for example we would be able to issue 40% of the sample initially to web and achieve a full household response in web of around 50%.

1. There is also a self-completion instrument for 10-15 year olds, but it isn’t relevant to this paper. [↑](#footnote-ref-1)
2. We randomly allocated households to a model and a test sample, but we needed the individual response rates at web and FTF to be similar in both halves. We had to draw several random half samples before we met the response rate criterion. [↑](#footnote-ref-2)
3. Suppose we were looking at selecting 10% of households to issue to web, we could apply the model to the FTF-issued sample, and then look at response among the corresponding 90% who would NOT have been issued to web. [↑](#footnote-ref-3)
4. Therefore in a 4-adult household, if 3 people took part at the previous wave then if all these 3 take part at the current wave, that would be defined as a fully productive household, regardless of the interview status of the 4th adult. [↑](#footnote-ref-4)
5. This assumes that the probabilities of people in one household taking part are independent, which they almost certainly aren’t. [↑](#footnote-ref-5)