# Survey Data Analysis Final Report

*Sanne Meijering*
*Hanne Oberman*
*Gerbrich Ferdinands*

## Introduction

The Innovation Panel is...

Multistage stratified cluster sampling was used...

Central to this research report are the following research questions: '...'.

In this report, we will first look at the sampling design and design weights. We will then calculate population estimates for some parameters. Finally, we will investigate the sources and effects of non-response errors.

## Questions

### Sampling Design

#### Question 1

For the Understanding Society study, up to three households per dwelling and up to three people per household were selected for interviews. If more than three households were found at a single dwelling, or if a single household contained more than three people, the Kish Grid method was used to select households or people at random.

> This method considered to lead to a random sample of household members, and avoids selection bias in the survey. (Why?)
>
> It is a popular method because when carried out correctly, it leads to (almost) equal probability sampling, something other selection methods do not obtain. Moreover, the only information you need in order to select respondents is a list containing the names of the persons in the household you want to sample. (why?)**

#### Question 2

Using the sampling design described above, people were assigned design weights. There is very little variance between weights (0.02), which is due to rescaling of the weights, as explained in the IP User Guide (p. 56): "Each set of weights has been scaled by a constant factor to produce a mean of one amongst cases eligible to receive the weight". Rescaling, truncating, or smoothing of weights are techniques to inhibit the effect of a single observation on population estimates >>>>>> (Lohr, **year**).

> Lower MSE (Lohr, p. 286), and application from IP User Guide.

#### Question 3

The post-stratified weight was calculated using the design weights, sex, age and government office region. It is known that age was split up in seven groups and the government office regions were split up in four, but which ones these were is unknown.

In order to find out, we first turned age into a numeric variable and created dummy variables for all government office regions, with Scotland being the baseline. Sorting the coefficients from lowest to highest yielded the following graph:

```
#Insert graph
```

It is not completely clear which values should be grouped together into four groups, but there are two groups of three regions and one group of two with near-identical weights. As can be seen on the map below, where the regions are numbered from the lowest to the highest weight, the grouping is rather strange.

The baseline and the other two areas with a weight close to zero are Scotland, London and the West Midlands. Scotland and London are prime candidates to be categories of their own, as Scotland has a strong regional identity and London is the capital of the country. From a data-driven perspective, combining the West Midlands with any region other than London or Scotland does not make sense, as the weight difference between those region and all other regions is relatively big. From a theory-driven perspective, we cannot think of any good reason to combine the West Midlands with London or Scotland: it is not filled with cities nor is it close enough to Scotland to assume a similar population.

When assuming that the regions with a near-identical weight are in the same group and Scotland and London are two seperate categories, only a few possible groupings remain:

1. South East, South West, and North East England are the third category, with Wales and the rest of England being the last category.

- This makes some sense theoretically: the three are the very north and south of England, with the remainder being in the middle. Weight-wise however, combining the West Midlands with the areas with far lower weights does not make sense.

2. The same as above, but with the West Midlands grouped with London.

- This is more logical weight-wise than the first option, but from a theoretical perspective it is strange to combine the city with a rural area.

3. The North West, West Midlands and East of England are the third category, with Wales and the rest of England being the last category.

- Weight-wise it is just as logical or illogical as the first option, but from a theoretical standpoint taking three areas in the middle seems to have less of a basis than taking the outer areas like is done in the first option.

4. The West Midlands alone being the third category, with the rest of England being the fourth.

- This fits weight-wise, but why the West Midlands should be taken seperately is a mystery. It is however less absurd than combining it with London: the West Midlands may have unique features that we do not know about, but we do know that it is not a capital like London.

In conclusion, the fourth option seems the most likely, as the grouping is logical from a data-driven perspective and while it does not quite make sense theoretically, there is no obvious reason to assume that it is incorrect. The best alternative is the first option, as it makes the most sense from a theoretical perspective.


Figure 1: pic_gov