Appendices

A Coding demographic categories

In this article we use data from three different sources of data: univariate census statistics, a sample of anonymised records (SARS) from the Census, and opinion survey data. Because we use multiple sources of data, we must make sure that our variable categories are comparable. In this appendix we describe the categories used for each of our seven post-stratification variables, and the operations necessary to reconcile different categorisations.

The coding of *gender* is dichotomous (male/female); this is the same across all data sources.

The coding of *age* differs across our sources of data. The SARS data uses the following categories: 16-19, 20-24, 25-29, 30-44, 45-59, 60-64, 65-69, 70-74, and 75+. All other sources record age as a continuous variable. Consequently, we adopted the SARS categories. For the purposes of post-stratification, we created an artificial 18-19 age category by (1) taking the 16-19 category, and multiplying by one-quarter, to create one artificial year; (2) taking the 20-24 category, and multiplying by one-fifth, to create one artificial year; (3) adding the sum of these two categories, and using this value. We are therefore assuming that the joint distributions involving 16 to 19 year olds are very similar to the joint distributions involving 18 to 19 year olds.

The coding of *education* is the most problematic. We adopt the following categories, which are used in the SARS data, but which are not used in the considerably more detailed univariate statistics and survey data:

1. Qualifications data missing

2. No qualifications

3. Level 1

4. Level 2

5. Level 3

6. Level 4/5

7. Other qualifications/level unknown

These levels are similar to International Standard Classification of Education (ISCED) levels. As such, Level 4/5 corresponds to post-secondary educational attainment; Level 3 to attainment at the end of secondary education, and Levels 1 and 2 to lower secondary or primary educational attainment. Specific educational outcomes were recoded on this basis.

The coding of *marital status* involves collapsing detailed information from SARS and from opinion data to the following dichotomy, for which information is available in the census univariate statistics:

- Married or re-married
- Single (never married), separated, divorced or widowed

The coding of *housing status* involves collapsing detailed information from the SARS and univariate census data to the following dichotomy, for which information is available in the public opinion survey data:

- Owns accommodation
- Rents accommodation

The coding of *social grade* relies on the National Readership Survey/Market Research Society social grades

1. Approximated social grade AB
2. Approximated social grade C1
3. Approximated social grade C2
4. Approximated social grade DE

Note that this refers to the social grade of the 'head of household' or 'household reference person' (HRP). For public opinion survey data, we have been able to recode information on occupation to the above categories, using (where appropriate) information on the occupation of the respondent's partner, or information on their student status.

Finally, the coding of *private sector occupation* involves a simple dichotomy between those.

- currently in private sector employment
- in public or voluntary sector employment, or unemployed

B Generating post-stratification weights

The UK Census Dissemination Unit provided 2001 census results at the Westminster parliamentary constituency (WPC) level. Three types of information are provided: univariate counts for all census variables at WPC level; bivariate cross-tabulations for a limited number of combinations of variables at WPC level; and a sample of anonymized records (SARS) which permits multivariate cross-tabulations for all census variables at national level. In this appendix we describe how we used the SARS together with the univariate statistics to estimate the joint distribution of our variables at WPC level, and how we checked these estimates against the available bivariate cross-tabulations.

We began with the SARS data, and created a six-dimensional matrix (2 genders $\times$ 9 age categories $\times$ 7 education categories $\times$ 2 marital statuses $\times$ 2 housing statuses $\times$ 4 social grades $\times$ two sectors of the economy (private and public)). Due to the changes in the education systems of England, Wales and Scotland over time, information on the educational attainment of over 75s was not included. We therefore estimated, using those respondents in the 65-74 age group only, a multinomial model of educational attainment using all of the remaining variables in our matrix as predictors. We used the predicted probabilities of attainment in each category to create estimated counts for each cell.

We then created as many copies of this six-dimensional matrix as there were WPCs (632). Call each of these the target matrix. For each constituency, and for each variable, we multiplied the entries in the target matrix by the proportion to which they were under-represented compared to the known marginal distribution provided by the Census Dissemination Unit. Thus, for Aberdeen North, the proportion of women in the population according to univariate census statistics (51.6%) is slightly lower than the proportion of women in the SARS (51.8%); and so all cells in the target matrix involving women were multiplied by 0.995, and all cells in the target matrix involving men were multiplied by 1.005. We finished this

iterative 'raking' process when the mean absolute logged difference in these proportions was less than 0.0001. The result was an estimate of the joint distribution of variables in each constituency based on the known national joint distribution of variables as adjusted for the over-/under-representation of certain groups in each constituency.

In order to verify that these raked estimates were reliable, we compared our estimates to the limited bivariate cross-tabulations made available at WPC level by the Census Dissemination Unit. Here, we use tables CAS033 (Occupation by age) and CAS113 (Occupation by educational attainment). We converted these counts to notional weights by dividing by the grand sum. We can assess the congruence between our estimates and the actual Census joint distribution by calculating the absolute difference in the weight for each cell, and averaging across constituencies. The mean absolute difference for CAS033 was 1.28%; the mean absolute difference for CAS113 was 0.24%.