# UMbRELLA interim report – Preparatory work

This document is intended to supplement the UMbRELLA Interim Report 2 (January 2016) by providing a summary of the preliminary analyses which influenced the decision not to apply a weighting adjustment to the data produced by the Doctors' Census survey.

Weighting adjustment is a procedure employed in survey research to minimise potential bias resulting from unrepresentative sampling of a wider population. The decision of whether or not to weight data is largely subjective, as are many of the decisions that need to be taken in calculating weightings (Biemer & Christ, 2008; Gelman, 2007). We present a brief summary of the theoretical and practical issues relevant to the decision not to weight data from the Doctors' Census survey.

The survey was disseminated to a predetermined mailing list of all eligible doctors registered and licensed with the GMC, but importantly doctors in training were excluded. Firstly, we compared survey respondents with non-responders as well as the mailing list as a whole across several key demographic variables (respondent characteristics) provided by the GMC. These were sex, age, region, place of primary medical qualification (PMQ), prescribed connection, speciality, and GP status. Overall our sample of respondents was broadly representative of the population as a whole, with only small differences revealed for age and ethnicity. Secondly, we conducted analyses on responses to survey questions of interest using both weighted and unweighted data to ensure that these small differences did not introduce bias in the conclusions drawn from the survey findings.

## 1 Whether to Weight Data

### 1.1 Non-Response Bias

Non-response bias can be either at the unit or item level. Unit non-response is when a number of individuals in a target population are unable to be contacted to take part in a survey or do not respond to a survey invitation. Item non-response refers to missing responses from individuals on single items within the survey (Lynn, 1996). Bias can potentially arise as a result of both types of non-response.

Unit non-response is more relevant to the generalisability of the conclusions reached based on the data. However, whether or not a particular group is more or less likely to respond does not automatically imply biased conclusions. It may only be an issue if those in an over- or under-represented group hold views that are different to other groups. For example, in estimating agreement with an item, if young and old respondents do not differ in their views, over- or under-representation of a given age-group will not bias the overall estimate of agreement (Lynn, 1996).

Statistically, the existence of bias due to unit non-response can be evaluated by comparing item means between responders and non-responders or by comparison of the probability of responding for each subgroup (Holt & Elliott, 1991). Output from the Doctors' Census survey does not include item data from non-responders; therefore we cannot say whether responders are representative of non-responders in their answers. This means we can only evaluate the representativeness of the sample based on its demographic make-up; any difference between responders and non-responders in their experience of and views on revalidation cannot be eliminated by weighting on demographic variables alone. The impact of any over- or under-representation relies on evaluation of any impact of weighting, and any effects of demographic factors on the item responses. If no differences between demographic subgroups are found in the item responses, weighting will not alter results. If differences are found, but

weighting has no material effect on conclusions, then the decision to not weight is justified statistically, although there are also methodological factors to consider.

### 1.2 Sample Representativeness

As UMbRELLA have demographic data for the entire population of interest (the mailing list), we compared the proportions in each subgroup between the population and those who responded to the survey. Based solely on this type of comparison, the sample characteristics are extremely close to those of the population (see Appendix A). The largest, though still small, differences are seen for age and ethnicity. Although the decision over whether a particular distribution is representative or not is largely subjective, particularly with a sample of this size, some attempts have been made to quantify representativeness of survey samples. One such approach is to use a logistic model to determine response probabilities for everyone in the mailing list, and use these to calculate an R-Indicator ($R_{rho}$, Bethlehem, Cobben, and Schouten, 2008) which ranges from 0-1, with 1 indicating 'strong representativeness'. The $R_{rho}$ value for the Doctors' Census survey is 0.868. It should be noted that a number of assumptions are required for such modelling, so although one can apply the technique to explore measures of representativeness, it cannot be used to calculate weightings in this case.

Predicting the probability of a survey response based on all demographic factors suggested that only the oldest age groups were significantly more likely to respond, and non-white ethnic groups were less likely to respond. Therefore older age groups are slightly over-represented in the sample, as are white doctors, but even in these cases the sample can be considered "broadly representative".

As mentioned above, over- or under-representation in the sample does not necessarily imply bias. Although we have no item data for non-responders, we can assess differences in responses to items between groups in the sample, and interactions between different demographic characteristics.

Given this and the fact that there is a degree of subjectivity in many methodological and analysis choices, we initially performed analyses on both weighted and unweighted data.

The choice of weighting method and how to implement it should play a role in the decision to weight or not. If the method for calculating weights rests on a number of indefensible assumptions then any conclusions drawn from that data are also undermined. If weighting takes into account some factors but not others then any conclusions need caveats which describe the particular factors that have been weighted for. After considering two methods, proportional weights were implemented with our survey data. Please see Appendix B for further discussion of the methods and rationale for our decisions.

## 2 The Effect of Weighting on Survey Analyses

We found an effect of sex, age, and ethnicity but that the interpretation of the analyses are unaffected by weighting. This would suggest that although older age-groups and white doctors might be marginally over-represented in the sample, after taking into account all respondent characteristics this does not affect the conclusions. We describe the details below:

Analyses of Yes-No responses by age-group to questions such as Q34, "Did you change any aspects of your clinical practice, professional behaviour, or learning activities after your recent appraisal?" show that the inferences drawn do not differ between the weighted and unweighted data. Both show the same associations between age-group and response to Q34 at the same level of significance; $X^2(6, n=23115)=382.28$, $p<0.001$ for unweighted data, $X^2(6, n=23115)=298.53$, $p<0.001$ for weighted data.

As additional example cases, regression analyses were conducted on weighted and unweighted data to assess the effect of sex on "difficulty of collecting supporting information" (Survey Q57.1), and age on agreement that "appraisals are an effective way of helping doctors improve their clinical practice" (Survey Q42.1). In both examples weighting has no impact on the conclusions drawn from the inferential statistics. The only effect is to adjust the proportions of each type of response – though this presents difficulties when interpreting results (see Section 4 below). Weighting was based on sex, age, ethnicity, region, PMQ region, prescribed connection, specialty group, and GP status (as described in Appendix B).

Using unweighted linear regression to determine the effect of sex on Q57.1 responses shows that males have an average response 0.163 points below females, $R^2_{adj} = 5.615x10^{-3}$, $F$ (1,23056) = 131.2, $p$ <0.001. The weighted equivalent of this estimate shows males have an average response 0.158 points below females, $R^2_{adj} = 5.346x10^{-3}$, $F$ (1,23056) = 123.9, $p$ <0.001.

It is important to note that the significance level is likely to be highly biased by the large sample size, and therefore the interpretation of effect size becomes more relevant in determining whether or not there is a meaningful difference between groups. Using Cohen's $d$ as a measure of effect size, the male-female difference has an effect size of $d$ = 0.153 and $d$ = 0.149 for unweighted and weighted data respectively; which are comparable effect sizes.

Using unweighted and weighted linear regression to determine the effect of age on responses to Q42.1 provides the coefficients and significance values shown in Table 1. Although the coefficients for the age groups 20-29, 30-39, and 70+ have different $p$-values when weighted data are used to fit the model, the coefficients remain comparable. Again, both analyses would lead to the same inferences.

**Table 1 Weighted and Unweighted regression coefficients and associated probability for Age**

| | Unweighted | | | Weighted | | |
|---|---|---|---|---|---|---|
| Age | Coeff. | *p* | Sig. | Coeff. | *p* | Sig. |
| <20 (ref) | --- | --- | --- | --- | --- | --- |
| 20-29 | 0.2314 | 0.014637 | <0.05 | 0.23366 | 0.00662 | <0.01 |
| 30-39 | 0.19279 | 0.000918 | <0.001 | 0.17396 | 0.01241 | <0.05 |
| 40-49 | 0.05335 | 0.346505 | ns | 0.07801 | 0.25872 | ns |
| 50-59 | -0.04918 | 0.384569 | ns | -0.01012 | 0.88434 | ns |
| 60-69 | 0.04894 | 0.402070 | ns | 0.09921 | 0.17104 | ns |
| 70+ | 0.20306 | 0.004366 | <0.01 | 0.19851 | 0.02888 | <0.05 |

Using unweighted and weighted linear regression to determine the effect of ethnicity on Q42.1 responses provide the coefficients and significance values shown in Table 2. Regardless of whether weighted or unweighted data are used, the coefficients and $p$-values are comparable. In this case both analyses also lead to the same inferences.

**Table 2 Weighted and Unweighted regression coefficients and associated probability for Ethnicity**

| | Unweighted | | | Weighted | | |
|---|---|---|---|---|---|---|
| Ethnicity | Coeff. | *p* | Sig. | Coeff. | *p* | Sig. |
| Asian (ref) | --- | --- | --- | --- | --- | --- |
| Black | 3.38165 | $< 2x10^{-16}$ | <<0.001 | 3.34454 | $< 2x10^{-16}$ | <<0.001 |
| Mixed | 0.27237 | $3.87x10^{-09}$ | <<0.001 | 0.2898 | $1.40x10^{-9}$ | <<0.001 |
| Other | -0.26156 | $2.26x10^{-06}$ | <0.001 | -0.20718 | 0.0004252 | <0.001 |
| Undeclared | -0.11329 | 0.0129 | <<0.05 | -0.13289 | $< 2x10^{-16}$ | <<0.001 |
| White | -0.60517 | $< 2x10^{-16}$ | <<0.001 | -0.61809 | $< 2x10^{-16}$ | <<0.001 |

In these examples, although effects of gender, age, and ethnicity are found, the interpretation of the analyses is unaffected by weighting. In relation to earlier points about potential bias, this would suggest that although older age-groups and white doctors might be marginally over-represented in the sample, after taking into account all respondent characteristics this does not affect the conclusions. Furthermore, exploratory ANOVA models (not shown here) which assess the effects and interactions of all respondent characteristics on a range of questions relating to appraisal and revalidation (all parts of Q42, Q69 and Q70), show no interactions between age and ethnicity at this stage. This suggests that views about appraisal and revalidation and their variation across respondent characteristics are much more nuanced than any clear dichotomous divides across demographic lines.

For completeness, exploratory work has been conducted comparing the output from the multifactorial ANOVAs with and without adjustments for weightings. Though the theory underpinning the inclusion of sample weights in complex analysis of variance designs lacks consensus, the results show that patterns of significance hold across both weighted and unweighted models.

## 3 Implications of conducting post-survey weight adjustments

The analyses, as presented in our interim report, show frequency counts and percentages based on the data in the sample. Adjusting these counts and percentages using any form of weighting adds a level of abstraction such that claims of the form “X% of Group Y respond Z” lose any direct relationship to the data. The figures that are reported in weighted tables are not actual numbers of responses within each category, but extrapolated estimates based on the particular form of weighting employed.

It is perhaps much more important to be able to say “X% of those who responded thought Y” than to present extrapolated proportions. Claims based on extrapolated proportions would be open to the criticism that they are not based directly on actual responses, and would vary depending on the weights employed, which in turn would vary by weighting method and the choices made in its implementation.

Because of the exponential increase in complexity and abstraction, not weighting leads to more meaningful conclusions as analyses progress to comparisons within and between subgroups of interest. At this level of analysis the slight differences in absolute proportions of respondents in each subgroup relative to the population become less of a concern (see previous section). Not weighting the data allows claims to be made based on the actual responses collected, and the analyses employed take into account the fact that they are conducted on a sample which although not reflecting the population in absolute terms does broadly reflect its strata.

Another practical issue related to decisions about performing weighting adjustments on survey data that must be considered is how to include accurately weighted data in the multivariate analyses.

With respect to implementation, where comparisons were made between weighted and unweighted data in order to determine the impact and necessity for weighting, weights were incorporated as 'sampling weights', not 'frequency weights'. The latter artificially inflates sample size and may bias any complex multivariate analyses. Sampling weights are only incorporated in some statistical packages, and how to use them accurately in complex analyses is open to debate (Wojtys, personal communication; see also Section 4). This means that weighting may be problematic due to a lack of theory for combining weightings with more complex analyses (Kish, 1990; Gelman, 2007). This may in turn leave conclusions based on weighted analyses open to criticisms on methodological and theoretical grounds – criticisms that could not be levelled at conclusions based on unweighted data.

# 4 Summary

The results of our initial comparison of survey respondents with non-respondents on the demographic data provided by the GMC show that although older age groups and white doctors are marginally over-represented in the sample, and more likely to have responded to the survey, the sample can be considered broadly representative of the population of interest.

However, opinions over what constitutes 'broadly representative' may differ, meaning that the decision to weight is potentially contentious. Our analyses using both weighted and unweighted data have demonstrated that post-sampling weighting based on all demographic factors does not lead to significantly different conclusions.

As discussed in Section 3, performing weighting adjustments adds an element of abstraction to the interpretation of the results, leading to difficulties in interpretation. Furthermore weighted data can lead to theoretical and practical constraints when performing multivariate analyses, which are of particular importance in understanding relationships between subgroups and their experiences of revalidation, a primary aim of the Doctors' Census survey.

Therefore, after careful consideration of the analyses presented above, as well as the implications of weighting on the interpretation of results and further analysis, the decision was made not to weight the data.

## Appendix A: Demographic profiles of responders, non-responders and the target population

**Table 3 Doctors (%) by sex and survey response**

| Sex | Responders (A) | Non Responders (B) | Mailing List (C) | A-B | A-C |
|---|---|---|---|---|---|
| Female | 41.36 | 41.86 | 41.77 | -0.5 | -0.41 |
| Male | 58.64 | 58.14 | 58.23 | 0.5 | 0.41 |
| Total | 100.00 | 100.00 | 100.00 | 0.0 | 0.00 |

**Table 4 Doctors (%) by age and survey response**

| AgeGroup | Responders (A) | Non Responders (B) | Mailing List (C) | A-B | A-C |
|---|---|---|---|---|---|
| Under25 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 |
| 25-29 | 2.13 | 4.51 | 4.11 | -2.38 | -1.98 |
| 30-34 | 6.22 | 10.97 | 10.18 | -4.75 | -3.96 |
| 35-39 | 12.40 | 18.61 | 17.57 | -6.21 | -5.17 |
| 40-44 | 14.90 | 19.45 | 18.69 | -4.55 | -3.79 |
| 45-49 | 15.22 | 15.54 | 15.49 | -0.32 | -0.27 |
| 50-54 | 16.42 | 12.73 | 13.35 | 3.69 | 3.07 |
| 55-59 | 15.38 | 9.05 | 10.11 | 6.33 | 5.27 |
| 60-64 | 9.19 | 4.96 | 5.67 | 4.23 | 3.52 |
| 65-69 | 5.19 | 2.64 | 3.07 | 2.55 | 2.12 |
| 70+ | 2.93 | 1.54 | 1.77 | 1.39 | 1.16 |
| Total | 100.00 | 100.00 | 100.00 | 0.00 | 0.00 |

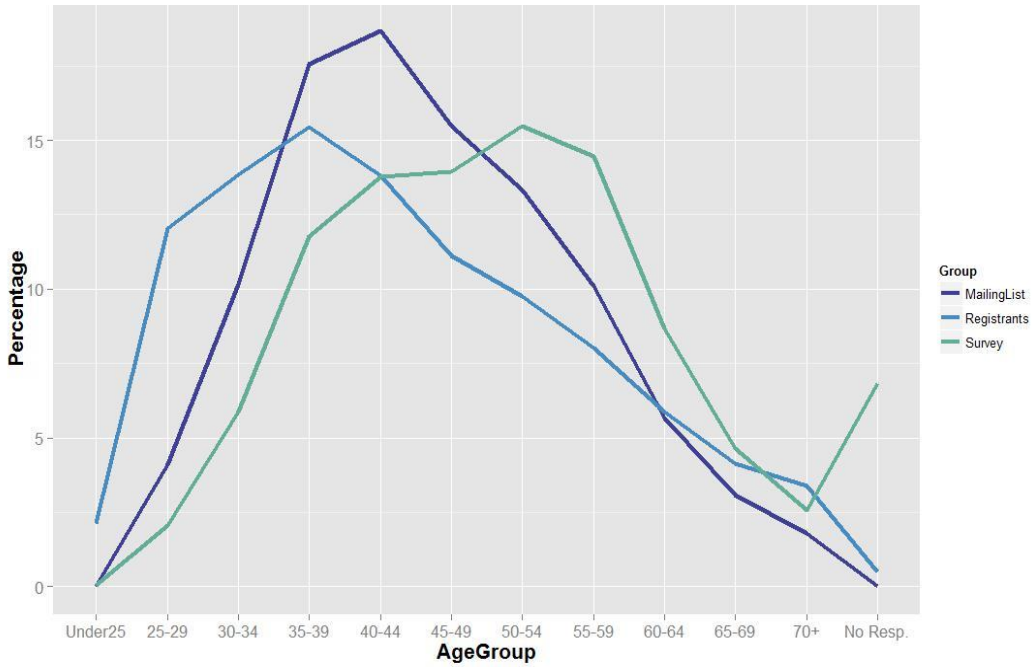


*Figure 1 Doctors (%) by Age group and survey response*

**Table 5 Doctors (%) by ethnicity and survey response**

| Ethnicity | Responders (A) | Non Responders (B) | Mailing List (C) | A-B | A-C |
|---|---|---|---|---|---|
| Asian or Asian British | 20.32 | 24.02 | 23.40 | -3.70 | -3.08 |
| Black or Black British | 3.20 | 3.43 | 3.39 | -0.23 | -0.19 |
| Missing | 15.29 | 17.62 | 17.23 | -2.33 | -1.94 |
| Mixed | 1.38 | 1.65 | 1.60 | -0.27 | -0.22 |
| Not stated | 0.50 | 0.88 | 0.82 | -0.38 | -0.32 |
| Other Ethnic Groups | 2.44 | 2.59 | 2.56 | -0.15 | -0.12 |
| White | 56.86 | 49.81 | 50.99 | 7.05 | 5.87 |
| Total | 100.00 | 100.00 | 100.00 | 0.00 | 0.00 |

**Table 6 Doctors (%) by region and survey response**

| Region | Responders (A) | Non Responders (B) | Mailing List (C) | A-B | A-C |
|---|---|---|---|---|---|
| Ch.Is, IoM, Overseas, Not Rec | 8.10 | 6.89 | 7.09 | 1.21 | 1.01 |
| East Midlands | 5.25 | 5.27 | 5.26 | -0.02 | -0.01 |
| East of England | 8.01 | 7.67 | 7.73 | 0.34 | 0.28 |
| London | 15.31 | 16.40 | 16.22 | -1.09 | -0.91 |
| North East | 3.41 | 3.61 | 3.57 | -0.20 | -0.16 |
| North West | 9.25 | 10.15 | 10.00 | -0.90 | -0.75 |
| Northern Ireland | 2.09 | 2.51 | 2.44 | -0.42 | -0.35 |
| Scotland | 8.34 | 8.01 | 8.07 | 0.33 | 0.27 |
| South East | 13.00 | 12.07 | 12.22 | 0.93 | 0.78 |
| South West | 7.52 | 7.16 | 7.22 | 0.36 | 0.30 |
| Wales | 4.48 | 3.99 | 4.08 | 0.49 | 0.40 |
| West Midlands | 7.65 | 7.73 | 7.72 | -0.08 | -0.07 |
| Yorkshire and The Humber | 7.59 | 8.54 | 8.38 | -0.95 | -0.79 |
| Total | 100.00 | 100.00 | 100.00 | 0.00 | 0.00 |

**Table 7 Doctors (%) by PMQ region and survey response**

| PMQRegion | Responders (A) | Non Responders (B) | Mailing List (C) | A-B | A-C |
|---|---|---|---|---|---|
| EEA | 14.39 | 12.13 | 12.51 | 2.26 | 1.88 |
| IMG | 28.85 | 29.70 | 29.56 | -0.85 | -0.71 |
| Not Recorded | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 |
| UK | 56.76 | 58.17 | 57.93 | -1.41 | -1.17 |
| Total | 100.00 | 100.00 | 100.00 | 0.00 | 0.00 |

**Table 8 Doctors (%) by prescribed connection and survey response**

| PresConn | Responders (A) | Non Responders (B) | Mailing List (C) | A-B | A-C |
|---|---|---|---|---|---|
| N | 4.64 | 4.22 | 4.29 | 0.42 | 0.35 |
| Y | 95.36 | 95.78 | 95.71 | -0.42 | -0.35 |
| Total | 100.00 | 100.00 | 100.00 | 0.00 | 0.00 |

**Table 9 Doctors (%) by speciality and survey response**

| Speciality | Responders (A) | Non Responders (B) | Mailing List (C) | A-B | A-C |
|---|---|---|---|---|---|
| N | 55.01 | 58.23 | 57.69 | -3.22 | -2.68 |
| Y | 44.99 | 41.77 | 42.31 | 3.22 | 2.68 |
| Total | 100.00 | 100.00 | 100.00 | 0.00 | 0.00 |

**Table 10 Doctors (%) by GP status and survey response**

| GP | Responders (A) | Non Responders (B) | Mailing List (C) | A-B | A-C |
|---|---|---|---|---|---|
| N | 68.1 | 67.11 | 67.27 | 0.99 | 0.83 |
| Y | 31.9 | 32.89 | 32.73 | -0.99 | -0.83 |
| Total | 100.0 | 100.00 | 100.00 | 0.00 | 0.00 |

## Appendix B: Technical Notes

As an exploration of its impact, the following sections detail the calculation of weights by two methods; response probabilities derived from a logistic regression model, and proportional weighting based on population versus sample characteristics (for further explanations of these methods, see Lynn, 1996, though see Holt & Elliot 1991).

### Weighting based on Logistic Regression

In the case of logistic regression weights, these are based on the predicted likelihood that an individual would respond given their demographic characteristics. Keeping all subgroups of all demographic variables violated a number of assumptions for logistic modelling, so a simplified model was considered whereby the significance and effect sizes for each predictor were used to exclude predictors. Given that this simplification would be based solely on statistical grounds, an alternative approach was adopted whereby weighting is proportional to the number of individuals in each possible subgroup (i.e. number of doctors with each possible combination of characteristics).

### Proportional Weighting

Proportional weighting requires the calculation of the number of doctors in each possible subgroup (i.e. number of doctors with each possible combination of characteristics) for the population and sample, and then uses the reciprocal of the response rate within each subgroup as the proportional weight for each individual in a given subgroup. This incorporates all levels of all factors and the relationships between them rather than weighting on a single factor or small subset, effectively weighting by the number of respondents in each category as a proportion of the number of doctors in that category on the mailing list.

Cross tabulations of all variables were used to determine the number of individuals in each sub-category. From this, the number of target respondents within each subcategory can be compared to the number of individuals within each category who responded to the survey. Response rates can then be calculated as (survey sub-category cell-count) / (mailing list sub-category cell-count). The reciprocal of this is then used to provide a weighting for the subcategory. Table 11 shows a sample of cases.

**Table 11: A sample of cases used for proportional weighting**

| Category Code | N (Mailing List) | N (Responses) | Response Rate | Weighting |
|---|---|---|---|---|
| 1.10.1.1.3.2.2.1 | 13 | 5 | 0.385 | 2.600 |
| 1.10.1.10.1.2.1.1 | 2 | 2 | 1.000 | 1.000 |
| 1.10.1.11.1.2.2.1 | 5 | 2 | 0.400 | 2.500 |
| 1.1.1.13.2.2.1.1 | 2 | 1 | 0.500 | 2.000 |
| 2.9.6.9.3.2.2.1 | 1 | 1 | 1.000 | 1.000 |
| 1.10.1.1.1.2.2.1 | 61 | 33 | 0.541 | 1.848 |

## References


Barnier, J., Briatte, F., and Larmarange, J. (2015) Questionr: Functions to Make Surveys Processing Easier. R package version 0.4.3. http://CRAN.R-project.org/package=questionr.

Bethlehem, J., Cobben, F., and Schouten, B. (2008) Indicators for the representativeness of survey response. Proceedings of the Statistics Canada symposium.

Biemer, P. P., & Christ, S. L. (2008). Weighting survey data. International handbook of survey methodology, 317-341.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (2009) Introduction to Meta-Analysis (Ch7). London; John Wiley & Sons.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. Statistical Science, 153-164.

Groves, R.M. and Peytcheva,E. (2008) The impact of nonresponse rates on nonresponse bias. Public Opinion Quarterly, 72(2) 167-189.

Harrell, F.E. with contributions from Dupont, C. and others (2015) Hmisc: Harrell Miscellaneous. R package version 3.16-0. http://CRAN.R-project.org/package=Hmisc.

Holt, D., & Elliot, D. (1991) Methods of weighting for unit non-response. The Statistician 40(3), 333-342.

Lumley, T. (2014) Survey: analysis of complex survey samples. R package version 3.30. http://CRAN.R-project.org/package=survey.

Lynn, P. (1996). Weighting for non-response. In R.Banks, J.Fairgrieve, L.Gerrard, T.Orchard, C.Payne & A.Westlake (Eds.), Survey and statistical computing 1996 (pp. 205-214). Essex, UK: Association for Survey Computing.

Kish, L. (1990) Weighting: Why, When, and How? In Proceedings of the Section on Survey Research Methods, American Statistical Association, (pp. 121-130).