Weighting Guidance

Understanding Society has a complex design. The dataset allows a vast array of research across different disciplines, topics and population sub-groups.

As well as providing a rich dataset for researchers, Understanding Society also has to consider the respondent burden of the survey and maintain a cost-effective data resource. Additionally, not everyone selected into the sample responds to the survey. As a result, the sample has a clustered and stratified design with certain types of people being over-represented in the sample (e.g., residents of Northern Ireland, ethnic minorities).

These features mean that it is important that researchers use weights in their analysis. This section of the user guide provides advice on the correct weight to use and the sample design variables (for clustering and stratification) provided with the data which will allow you to do this.

Why use weights?

Understanding Society is designed to be used with weights

Survey data, like Understanding Society, comes from a sample of a population. Weights appear in the dataset as a variable which assigns a value to each case to indicate how much 'weight' it should have during analysis. Using weights in your analysis means that your results will accurately reflect the population. Typically you will specify the weight before you start your analysis.

It's important to weight the data to represent and reflect the views and attitudes of the whole population; to hear from everyone across the country.

The benefits of using weights in your analysis means you can correctly reflect the population structure, as some groups may be over-represented in the sample. In Understanding Society we over sample ethnic minority respondents in the Ethnic Minority Boost Sample and use weights to allow for those respondents who don't respond to the survey - certain types of people are more likely to respond than others. If a question is asked to groups which are over or under represented in the sample, then responses will be biased for that group and we use weights to correct this.

Example *- if the wages of some ethnic minority groups are lower than the white British population and as there is a higher proportion of ethnic minorities in this sample than in the population, unweighted estimates of UK wages will be downwardly biased. In other words, use weights when running your analysis or it will not correctly reflect the population structure.*

A weighted analysis will adjust for the higher sampling fraction in Northern Ireland and for different probabilities of selection in the EMB and IEMB samples, as well as for response rate differences between subgroups of the sample. An unweighted analysis does not correctly reflect the population structure unless the assumptions below are true. It is suggested that researchers publishing or presenting unweighted estimates make these assumptions explicit.

If no weighting is used, an analysis of Understanding Society data assumes that all estimated parameters (means, measure of dispersion, model coefficients, etc) do not differ between:

- Northern Ireland and the rest of the UK
- people of ethnic minority origin and people of white British origin
- recent immigrants to UK and people who have stayed in the country longer
- people who live at an address with more than three dwellings or more than three households and those who don't
- people who responded at Wave 1 and those who did not
- people who continued to respond at later waves and those who did not
- people who responded to each particular instrument used in the analysis (individual interview, self-completion questionnaire etc.) and those who did not, see Lynn, Burton et al. (2012)

An unweighted analysis of the former-BHPS sample assumes that estimated parameters do not differ between:

- each of England, Scotland, Wales and Northern Ireland
- people who live at an address with more than three dwellings or more than three households and those who don't
- people who responded at Wave 2 of Understanding Society in 2010 and those who may have become non-respondents at any time since Wave 1 of BHPS in 1991
- people who keep responding in later waves of Understanding Society and those who stopped responding at any point of time between 1991 and the last year in your analysis

We therefore strongly suggest conducting weighted analyses of the *Understanding Society* data.

Weights are constructed by combining (i) design weights which adjust for unequal selection or sampling fraction and (ii) non-response weights which adjust for differential non-response and attrition at various stages (household level, within household at individual level, whether adult respondent completed self-completion questionnaire or not).

Selecting the correct weight

Given that Understanding Society is complex and multi-purpose we provide multiple weights to meet the different needs of users.

Separate sets of weights are provided for the different sets of samples:

- combined GPS and EMBS from wave 1
- former BHPS sample from wave 2
- combined GPS, EMBS and BHPS from wave 2
- combined GPS, EMBS, BHPS and IEMBS from wave 6

The sets of weights are not the same for all combined samples, reflecting the differences in their data collection process.

To get started watch our short video about selecting weights in Understanding Society. The video shows you how to select a weight by exploring the naming conventions. The name of each weight reflects the wave it represents, along with the questionnaire, the data file and the sample used, either within a wave (cross-sectional) or across waves (longitudinally). The rules are described below:

Cross-sectional or longitudinal analysis

If your analysis uses only data from one Wave, select the "xw" (cross-sectional) version of the weight. This weight is defined for all sample members who responded to the relevant questionnaire at a particular Wave. If your analysis uses data from multiple waves select an appropriate "lw" (longitudinal) version of the weight.

Example *- if your analysis only uses data from Wave 4, select the "xw" (cross-sectional) version of the weight. Wave 4 variables begin with* ***d_****. If your analysis uses data from multiple waves select an appropriate "lw" (longitudinal) version of the weight, e.g. If you are looking at waves 4 to 9, use the last wave in you analysis; wave 9 variables begin with* ***i_****.*

Hierarchy of analysis levels

For individual level analysis you may want to combine information from different questionnaires. In this situation please select the weight suitable for the lowest level according to the hierarchy table below:

Example *- if you are analysing cross-sectional data from Wave 1, using questions from both the proxy/full interview as well as from the self-completion questionnaire, then the correct weight will be* ***a_indscus_xw*** *– the weight for the self-completion questionnaire is level 1 in the table and is lower than the proxy/full interview questionnaire at level 3.*

Variable weight **a_indscus_xw** is designed for participants from Wave 1 **(a_)**, aged 16+ **(ind)**, who answered via the self-completion questionnaire **(sc)** from the general population and ethnic minority samples **(us)** and analysed within one wave **(xw)**.

Levels of analysis	**Questions available for**
4	Household level (all enumerated individuals)
3	Adult proxy and main interview
2	Adult main interview only (no proxy)
1	Adult or youth self-completion interview

After you have decided the population you want to generalise your results to, the sub-sample you want to use and the questionnaire(s) you want to use, refer to the following table which summarises the naming convention and makes it easy to choose the correct weight.

Table 43: Naming convention for Understanding Society weights

*w_*	**Xxx**	**Yy**	**Zz**	**_aa**
a_ **b_** **c_** **d_** **...**	**hhd:** household **psn:** persons 0+ **ind:** persons 16+ **yth:** persons 10-15	**en:** enumeration **in:** interview **px:** interview or proxy **5m:** "extra 5 minutes" **sc:** self-completion **ns:** nurse visit **bd:** blood	**us:** GPS & EMB **bh:** BHPS **ub:** GPS, EMB & BHPS **ui:** GPS, EMB, BHPS & IEMB **91:** BHPS original sample **01:** BHPS original sample + boosts	**_xw:** cross-sectional analysis weight **_lw:** longitudinal weight **_xd:** x-sectional design weight **_li:** longitudinal inclusion weight

* "gp" letters are used for weights available for the GP sample only. But there is only type of such weight - the design weights for the GP sample. This weight should be used by advanced users only.

For advanced users who want to model nonresponse in their own way, we provide design weights and inclusion weights which adjust the sample for unequal selection probabilities but not for nonresponse. Note that adjusting for the first wave nonresponse is different from adjusting for attrition and requires variables which have values for both responding households and never responding households.

How to use weights

In most statistical software packages, weights can be easily incorporated into most common forms of analysis (for example, using the svy commands in Stata or the Complex Samples module in IBM SPSS: see West et al, 2018, for a review).

Each Understanding Society weight is set to zero for all sample units to which it does not apply. Thus, specifying the use of the correct weight in analysis will automatically result in the analysis being restricted to the appropriate sample. For example, there are 39,289 persons in the file h_indresp, but only 27,841 of these have a non-zero value of h_indinui_lw. These are the people who gave a full individual interview at all of waves 6, 7 and 8. For longitudinal analysis of data obtained in the individual interviews at waves 6, 7 and 8, it is therefore sufficient to specify use of the weight h_indinui_lw. The analysis sample can of course also be further restricted by selecting based on respondent characteristics (e.g. by gender, or working status): the weight is appropriate for analysis of any demographic subset of the full sample to which the weight applies.

The weights provided have been developed for use when analysing data from various combinations of survey instruments in one of two ways:

When using data from a series of consecutive waves, e.g. a panel analysis. These are the longitudinal weights ending in **_lw**;

When using data from a single wave. These are the cross-sectional weights ending in **_xw**.

However given the richness of the data there are many other ways in which the data can be used. For example, you may want to use data from a non-consecutive set of waves, or for a calendar year, or relating to households, couples or other groupings in which each group member has participated. It is not always obvious how best to use the weights in these situations, so we have prepared a set of Frequently Asked Questions that address many of the most common uses of the study data. If your question remains unanswered, we also offer various forms of User Support.

Reference

West, B.T,, Sakshaug, J.W. and Aurelien, G.A. (2018) Accounting for complex sampling in survey estimation: a review of current software tools, Journal of Official Statistics 34(3): 721-752. doi: 10.2478/JOS-2018-0034.

Analysis advice for mixed mode data

Using different modes during a survey can affect how respondents' answer the same questionnaire.

Despite this possibility, the convenience and potential cost savings (especially relative to face-to-face/CAPI mode) have led Understanding Society to adopt a push-to-web mixed-mode design, starting at Wave 8, when 40% of participants were initially invited to complete the questionnaire online and a further 40% were initially approached for a face-to-face interview but then given the opportunity to complete online if they had not completed the face-to-face interview. The remaining 20% were only approached for a face-to-face interview. The implication of mode effects from Wave 8 onwards is that some of those people who chose web mode may have provided different answers to the same questions had they instead chosen CAPI. Given that 29% of Wave 8 individual interviews were carried out online, this means the introduction of mixed-modes could affect longitudinal analyses involving data from Wave 8 and earlier, predominantly CAPI, waves.

It is important to recognise that a substantively significant difference between the answers under web and under CAPI does not automatically imply that the web answer is 'worse'. CAPI is only a benchmark for comparison with data from earlier CAPI-mode waves. D'Ardenne, Collins et al. (2017) discuss how mode effects depend on several features of how respondents answer survey questions (fear of disclosure, social desirability bias for sensitive questions and positivity bias, satisficing), and the presentation of the question and its possible answers, so which mode is 'best' will depend on the nature of each question.

Wave 8 involved an experiment in which a proportion of households in the first year were randomized to receive web first or CAPI first. The data from this experiment allow the estimation of the effect of web mode on key statistics in a way that takes into account that within the experimental sample the characteristics of those responding online and those responding by CAPI may differ.

We are currently investigating issues for users and will provide more detailed advice in due course. Unfortunately, it was not possible to devise a simple fix to adjust the results of every longitudinal analysis to equal what would have been obtained had those choosing web counterfactually chosen CAPI. Instead, we offer the following advice for those users who wish to investigate the impact of web mode on their analyses:

1. **Do not use the 'indicator method' for a regression/multivariable analysis**: The indicator method is simply to include a dummy variable that indicates whether the user chose web or CAPI as predictor variable in the regression analysis. However, despite its popularity, it was found that this approach is generally ineffective because it can often lead to badly biased results.

2. **A simple sensitivity analysis is to compare the estimates obtained using only the ring-fenced sample with those obtained using the remaining data**: The ring-fenced sample is a random sample of 20% of households for which the survey was administered CAPI-only, as in previous waves. The variable **h_ringfence** identifies members of this sample. To test whether the results of a regression analysis are different in the ring-fenced sample from those in the mixed modes sample, the analyst can 1) include **h_ringfence** as a main effect in

the model, and 2) include the interactions between **h_ringfence** and each predictor variable in the model. We recommend that the survey design and weights are accounted for when performing this analysis. If any of the interactions created in step 2 are statistically significant, this indicates the potential presence of mode effects. If the results are significant and you are unsure of how to proceed, it is recommended that you consult a statistician on your team to discuss.

Results of the experimental analysis for future waves will be added when ready.

Clustering and stratification

As the sample design involves stratification, clustering and weighting, these design features affect standard errors and should therefore be taken into account in analysis. Appropriate variables are provided to allow the analyst to do this. Here we describe the stratification and clustering variables in the main Understanding Society data files. See Fumagalli, Knies et al. (2017) for a description of these variables in the harmonised BHPS. General advice on using this information appropriately applies to both the UKHLS and harmonised BHPS data.

The variable indicating the primary sampling unit is **psu.** It is available in the cross-wave files, **xwavedat xwaveid**. As the PSU is determined at the time of sampling the value of this variable does not change over time. But to make it easier to use, this variable is also included in wave specific data files, where the name of the variable is ***w*_psu** with "w_" reflecting the wave prefix. Similarly, as stratification occurs at the sampling stage, the variable representing stratification, **strata**, does not change over time and is available in the cross-wave files, **xwavedat xwaveid**, and in wave specific data files with a wave prefix, **w_strata.**