EXAMPLE 9:
USING THE BHPS COMPONENT OF UNDERSTANDING SOCIETY

EXAMPLE RESEARCH QUESTION(S): What is the co-evolution of subjective well-being and employment status in the United Kingdom over the last two decades, i.e. 1991 to 2010/11?

DESCRIPTION: Before Understanding Society: the UK Household Longitudinal Study (UKHLS) came into existence, there already existed a highly renowned longitudinal household panel study for the UK, the British Household Panel Survey (BHPS). From Wave 2 onwards the BHPS is part of Understanding Society. This example shows analysts how they can use the BHPS sample component within Understanding Society to create a long run of combined BHPS and UKHLS data. The first part of the worksheet introduces the BHPS data and documentation, then draws on techniques presented in Worksheets 2-6 to pull in and merge data from multiple waves, in this case using **`foreach`** loops. The second part introduces the UKHLS data and online documentation and provides some hints for combining UKHLS and BHPS data.

FILES: b_indresp (UKHLS); **aindresp**-**rindresp** (BHPS)
WAVES: Wave 2 (UKHLS); Waves 1-18 (BHPS)

STEPS:

0. Introduction
1. Use BHPS documentation to identify indicators.
2. Use Understanding Society documentation to identify indicators.
3. Construct BHPS data file from **windresp.dta** (BHPS W1-18)
4. Construct UKHLS data file from **w_indresp.dta** (UKHLS W2)
5. Append data created in step (4.) to data created in step (3.)
6. Analysis
7. Further tips for using UKHLS/BHPS data

NEW COMMANDS:

No new commands

9.0 INTRODUCTION

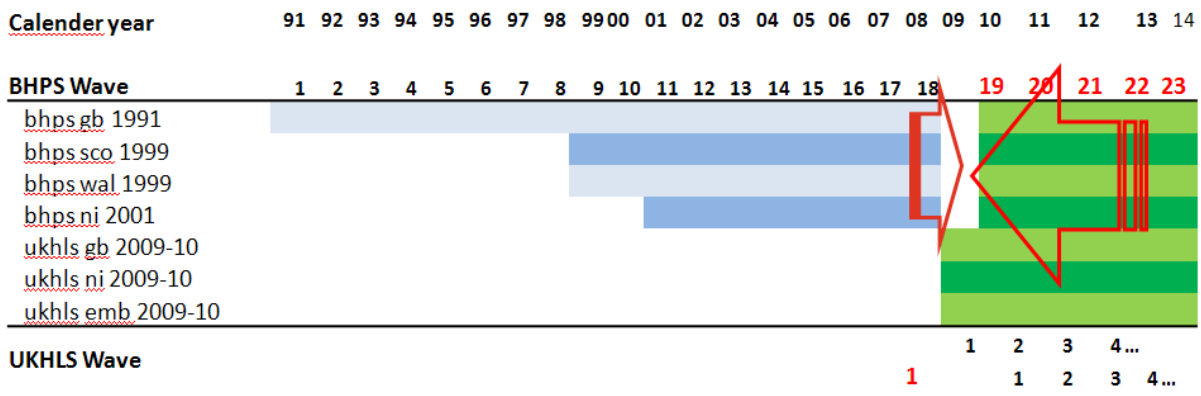
Understanding Society: the UK Household Longitudinal Study (UKHLS) draws on the success of her predecessor, the British Household Panel Survey (BHPS). BHPS is a longitudinal social survey of households and individuals living in the UK. It started in 1991 with 5,000 households selected at random within Great Britain. In 1999, an additional 1,500 households were added in each of Scotland and Wales and in 2000, an additional 2,000 households were added in Northern Ireland. Like Understanding Society, BHPS is a general topic social survey and covers areas such as demographics, household composition, employment, education, training, health, values and opinions and finances. See www.iser.essex.ac.uk/bhps for further information about the design and conduct of the study.

Data collection under the study title "British Household Panel Survey (BHPS)" stopped in 2008, i.e., in the year before Understanding Society started. The complete BHPS, Waves 1-18, is available for download at the UK Data Service (UKDS), see http://discover.ukdataservice.ac.uk/series/?sn=200005.

One of the decisions made with respect to the design of Understanding Society was to continue data collection from the BHPS sample members as part of the new study, and also to keep the design of the new study similar to that of BHPS. To this end, many of the questions planned to feature on Understanding Society were already carried on the last wave of the BHPS (i.e., Wave 18). Moreover, the Understanding Society questionnaire contains many questions that were previously carried on the BHPS.

At the time Understanding Society started, interviews with BHPS sample members for Wave 18 had only just concluded. So as not to overburden respondents with two interviews within 12 months, incorporation of the BHPS sample was aligned with data collection for Understanding Society Wave 2, with interviews being issued in the period January-December 2010 (i.e., year 1 of wave 2 only). From Wave 2 onwards, the UKHLS data package, available at http://discover.ukdataservice.ac.uk/series/?sn=2000053, includes data for the BHPS sample members. It is very easy to identify the BHPS sample members in Understanding Society as the unique cross-wave person identifiers in the BHPS, **pid,** are provided with each UKHLS data file (in Wave 2). It is also planned to re-release BHPS Waves 1-18, now with a unique Understanding Society **pidp** for each person that has ever participated in BHPS. This will ease linking files longitudinally across studies.

Analysts may treat information collected from BHPS sample members in Understanding Society Wave 2 as if it were information collected in BHPS Wave 19 or they may treat the BHPS Wave 18 data as if they had been collected as part of Understanding Society Wave 1.



In this example we briefly outline one possible use, i.e., treating the Understanding Society Wave 2 data for the BHPS sample as if it were Wave 19 of the BHPS.

To link information from Understanding Society and BHPS, users will need to be familiar with both studies. The studies' online documentations will be useful tools.

9.1 USING THE BHPS DOCUMENTATION TO IDENTIFY INDICATORS

In this worksheet we want to investigate the co-evolution of subjective well-being and employment over two decades, i.e., from 1991 to 2010. To this end, you need to identify indicators of subjective well-being that are available in Wave 2 of Understanding Society and in all waves of the BHPS; you also need an employment status indicator, and you need to pick appropriate weights provided in the study.

Due to time constraints we cannot provide you with a detailed overview of the BHPS. Please see the materials of our "Introduction to BHPS using Stata" course for a much more detailed introduction to BHPS. They are available at https://www.iser.essex.ac.uk/bhps/courses.
For the current example, Worksheets 0-3 and Worksheet 10 would be particularly useful.

The fact that the study designs of Understanding Society and BHPS are so similar comes in handy for analysts who know one of the datasets already. Broadly speaking BHPS has the same file structure as Understanding Society, and as a rule of thumb, if an Understanding Society variable already existed on BHPS it will have the same variable name. This is the case, for instance, for the employment status indicator **w_jbstat** on data file **w_indresp.dta**.

Similarities between the two datasets:

- One set of files for each wave, same root name, wave identified by a wave prefix
- Content of files with same names across the two datasets are the same. For example, **indresp** files always contain information collected from adult interviews, **hhresp** always contain information collected during household interviews and so on
- Variables also have the same root name across waves, with a wave prefix to identify the wave collected in
- Unique cross-wave identifier available
- Unique within-wave household identifier available. Note, again no concept of a longitudinal household

Despite all similarities there are a couple of noteworthy differences in the BHPS:

- Wave prefix is **w** (UKHLS: w_)
- Cross-wave person identifier: **pid** (UKHLS: **pidp**)
- Within-wave household identifier: **whid (**UKHLS: **w_hidp)**
- The postscripts **_dv** and **_xw** and **_lw** for derived variables and weights do not exist in the BHPS

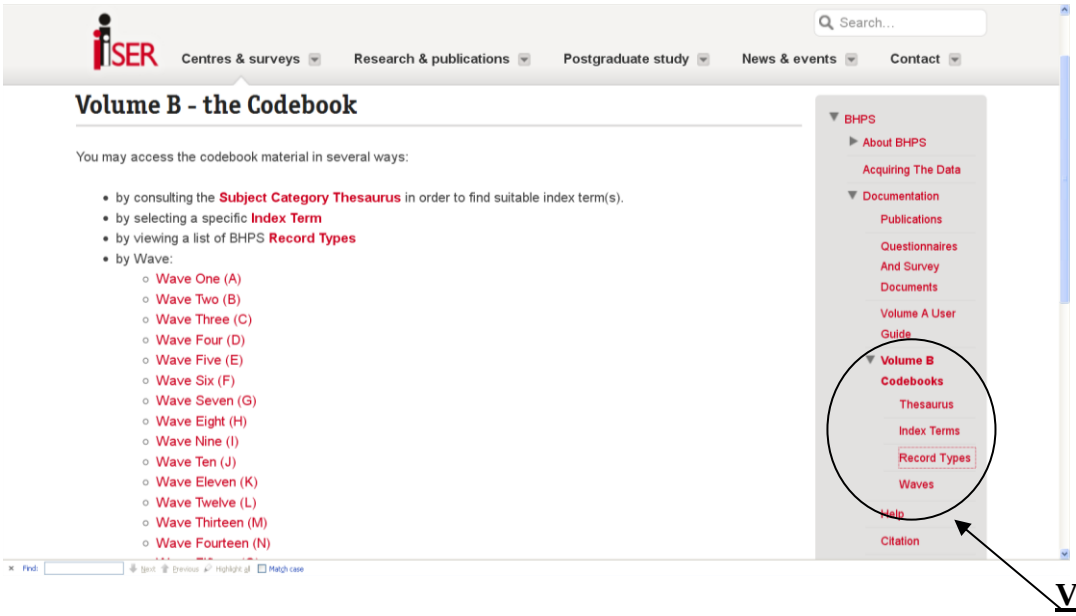
The BHPS documentation (including the questionnaires) is an essential tool, and available on our webpage: http://www.iser.essex.ac.uk/bhps

The online documentation covers (you can find links on the right panel of the webpage):

- Information on how to acquire the data (http://www.iser.essex.ac.uk/bhps/acquiring-the-data)
- Information on the sample (http://www.iser.essex.ac.uk/bhps/about/sample)
- Information on content of the questionnaire (http://www.iser.essex.ac.uk/bhps/about/questionnaire-content)
- Frequently Asked Questions (http://www.iser.essex.ac.uk/bhps/faqs)

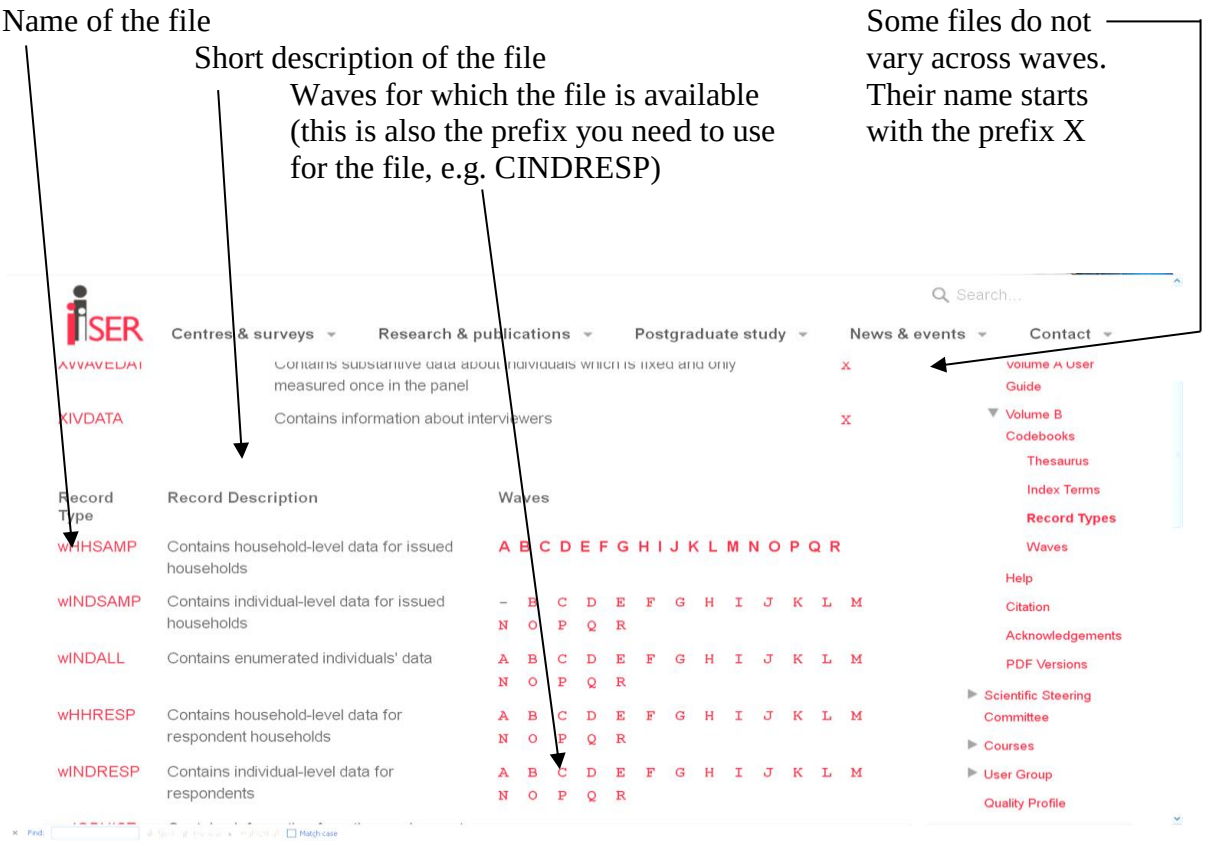
The Volume A of the User Guide is a pdf file which, among other things, includes descriptive information about survey and data:
http://www.iser.essex.ac.uk/bhps/documentation/vola/vola.html
In Appendix 2 you can find descriptions of the derived variables
In Appendix 3 you can find long coding frames

The Volume B of the User Guide is all online and should help you to identify which variables are available, where, for which waves, and so on. You can search Volume B either by subject thesaurus, index terms, record types or waves.



Volume B

To find out the names of the BHPS files we can start by following the Record Types link. In this page you can find the list, names, description, and availability of the different files.



In this exercise we will only use the INDRESP files. These are available for each wave.

You can find more information on each specific file by clicking on the name of the file. You can find the list of variables available in that file for a particular wave by clicking on the letters (e.g. 'C') on the right.

Task: Look at the BHPS online documentation and identify indicators of subjective well-being. Check availability across all waves of the BHPS!

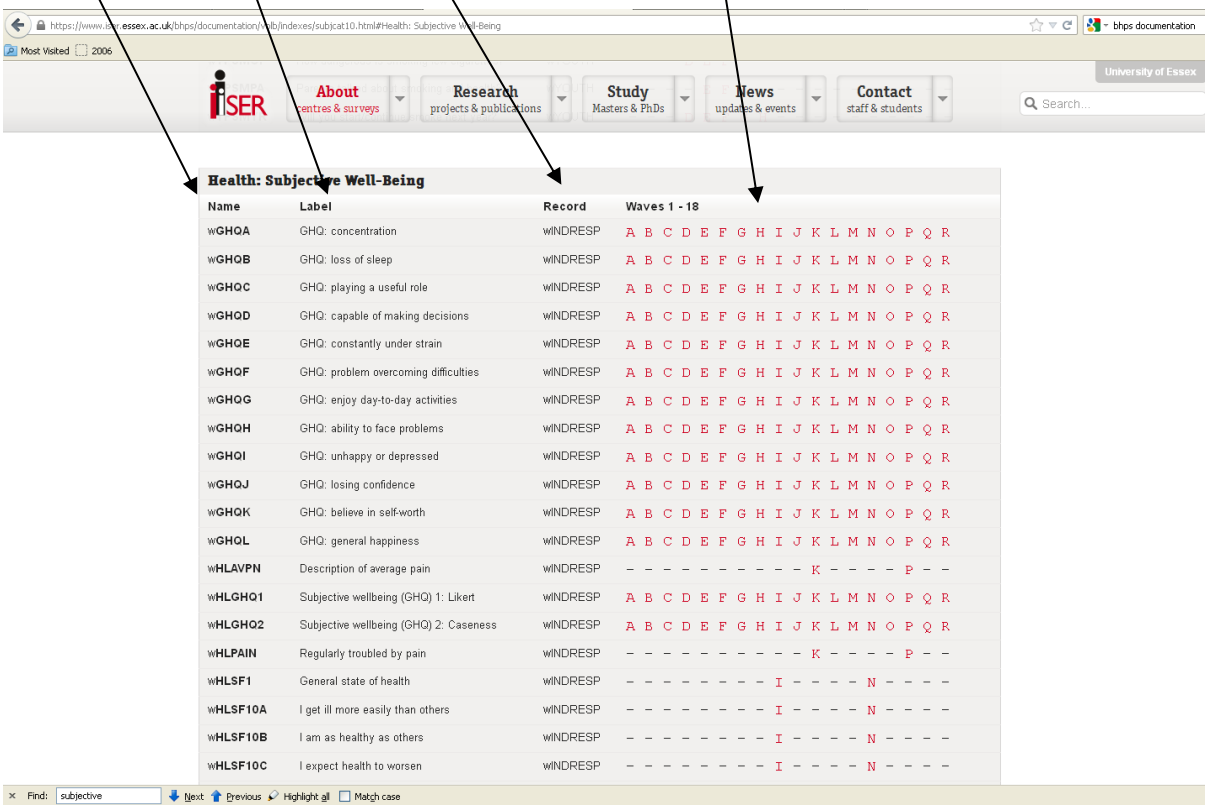
To find out if there is a variable identifying whether there are subjective well-being indicators we can start by consulting the Thesaurus page. There we can browse the terms until we find, for example: **Health: Subjective Well-Being**. If we click on the link, we arrive at a page showing:
(NOTE that we can arrive to this exact same page consulting the Index Terms web-pages)

Name of the variable ('w' means wave: 'a' for wave 1; 'b' for wave 2 and so on)

Label/description of the variable

File in which you can find the variable

Waves for which the variable is available

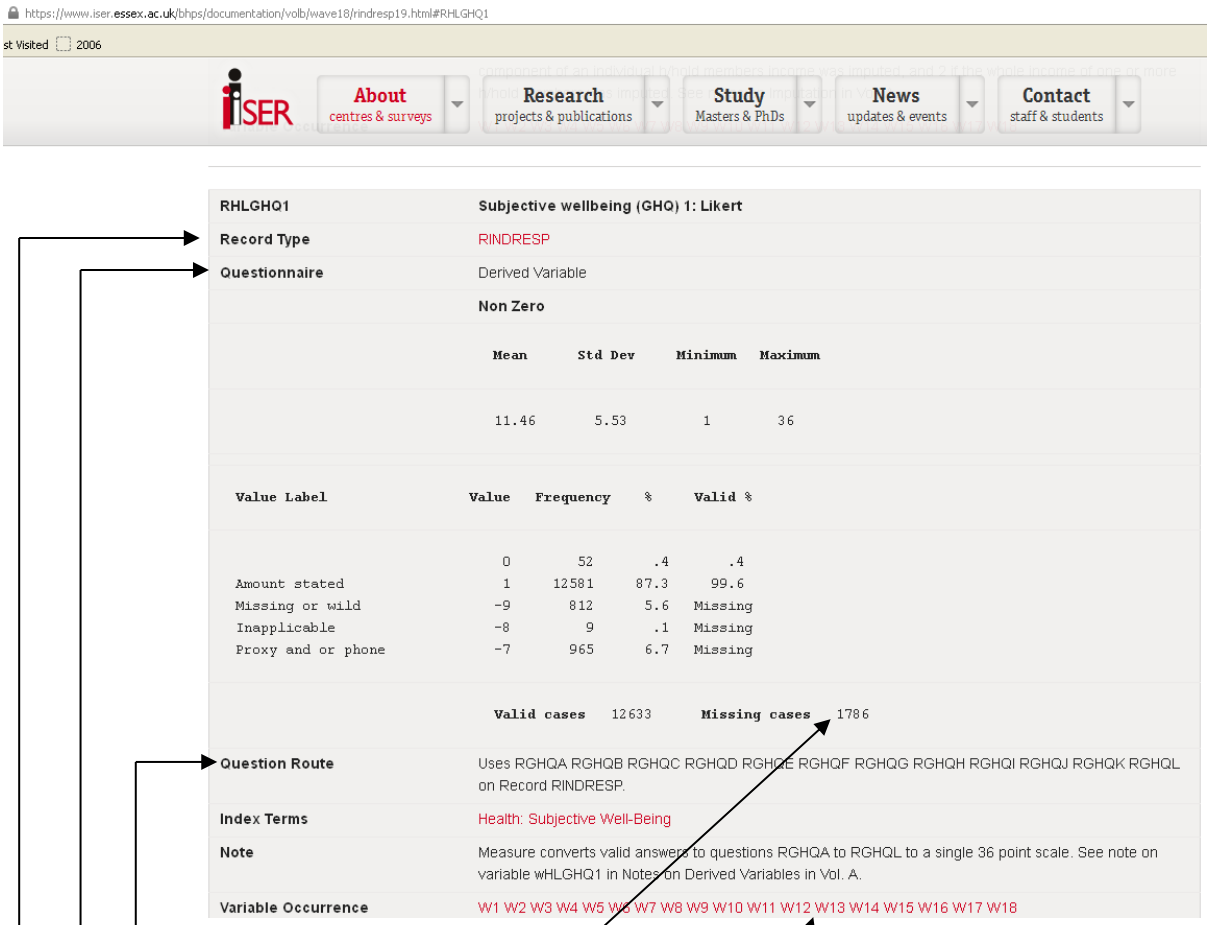


Health: Subjective Well-Being

Name	Label	Record	Waves 1 - 18
wGHQA	GHQ: concentration	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQB	GHQ: loss of sleep	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQC	GHQ: playing a useful role	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQD	GHQ: capable of making decisions	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQE	GHQ: constantly under strain	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQF	GHQ: problem overcoming difficulties	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQG	GHQ: enjoy day-to-day activities	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQH	GHQ: ability to face problems	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQI	GHQ: unhappy or depressed	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQJ	GHQ: losing confidence	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQK	GHQ: believe in self-worth	wINDRESP	A B C D E F G H I J K L M N O P Q R
wGHQL	GHQ: general happiness	wINDRESP	A B C D E F G H I J K L M N O P Q R
wHLAVPN	Description of average pain	wINDRESP	– – – – – – – – – K – – – – P – –
wHLGHQ1	Subjective wellbeing (GHQ) 1: Likert	wINDRESP	A B C D E F G H I J K L M N O P Q R
wHLGHQ2	Subjective wellbeing (GHQ) 2: Caseness	wINDRESP	A B C D E F G H I J K L M N O P Q R
wHLPAIN	Regularly troubled by pain	wINDRESP	– – – – – – – – – K – – – – P – –
wHLSF1	General state of health	wINDRESP	– – – – – – – – I – – – – N – – – –
wHLSF10A	I get ill more easily than others	wINDRESP	– – – – – – – – I – – – – N – – – –
wHLSF10B	I am as healthy as others	wINDRESP	– – – – – – – – I – – – – N – – – –
wHLSF10C	I expect health to worsen	wINDRESP	– – – – – – – – I – – – – N – – – –

*Question: There are many indicators of subjective well-being in BHPS. Why will this project draw on the General Health Questionnaire marker of subjective well-being (*w**hlghq1;** w**hlghq2**), and not the marker of life satisfaction (*w**lfsato**)?*

We can get more information on each variable by clicking on one of the waves, for example "R". For variable **rhlghq1** the following information is displayed:



*Task: We have also identified w**jbstat** as an appropriate indicator of employment status for our project. Browse the variable level information on w**jbstat** and carefully read the variable note at the bottom of the view!*

Hints: You could use the index term "Employment: Labour Force Status" to see which indicators mark employment status. However, you will remember from your work with Understanding Society that the variable is stored in the INDRESP data files. So another option would be to browse the INDRESP record of a specific wave and Ctrl+F to look for 'jbstat', or "employ" if you cannot remember the variable name, then click on the variable name.

It is now time to identify the appropriate weights for your analysis. To browse the weights follow the 'Index Terms' link → scroll down to and select 'Sampling Factors**'.**

Just like in Understanding Society, there are naming conventions for weights in BHPS, only they are not identical across studies. In BHPS, the first letter marks the wave, followed by 'x' or 'l' for cross-sectional or longitudinal, followed by 'r' for responding or 'e' for enumerated weight. There are separate weights for the original samples (variable names ending on 'wght') and for the extension samples (variable names ending on 'wtuk1' or 'wtuk2').

Naming convention for weights on BHPS

w	X	R	WGHT	Cross-sectional respondent weight in wave w, original sample
		E		Cross-sectional enumerated weight in wave w, original sample
		H		Cross-sectional household weight in wave w, original sample
	L	R		Longitudinal respondent weight in wave w, original sample
		E		Longitudinal enumerated weight in wave w, original sample
			WTUK1	Original sample + the extension samples
			WTUK2	Only extension samples separately

For a detailed discussion of the different samples and weights see Volume A, V.2.5. Seeing as we are interested in looking at cross-sectional data for GB (from 1991) and UK (from 2001) we will use w**xrwght** and w**xrwtuk1.**

Having identified which BHPS variables we are interested in, it is time to check the Understanding Society online documentation to see whether the indicators of interest are also available in that study.

9.2 USING THE UNDERSTANDING SOCIETY DOCUMENTATION TO IDENTIFY INDICATORS

*Task: Start by browsing the Understanding Society online **Dataset Documentation**, see https://www .understandingsociety.ac.uk/documentation/mainstage/dataset-documentation . Hint: You may want to take a quick look at Example 1 again.*

Task: Go to the datafile INDRESP and browse Wave 2. Using the Variable search function at the top left search for all variables you are interested in using the BHPS variable name stem!

Question: Does the Understanding Society Wave 2 data provide the employment status indicator? Can you confirm that the coding frame of this variable is identical to that on the BHPS Wave 2-18?

Question: Does the Understanding Society Wave 2 contain the Likert and caseness markers of GHQ?

If you have used the BHPS variable stem **hlghq1** or **hlghq2** as search terms, you will conclude that the information was not collected in Understanding Society. If you have searched for "GHQ" you will have concluded that the GHQ variables w**hlghq1** and w**hlghq2** exist in Understanding Society but are named **w_scgqh1_dv** and **w_scgqh2_dv**, respectively. There is, unfortunately, no general rule to make sure not to overlook a variable that existed in the BHPS. We advise to check the Understanding Society questionnaires carefully before abandoning a research idea. Searching for terms that are likely to feature in the variables label or description may also work better than searching for variable names. In future, the Understanding Society online documentation is likely to incorporate links to similar or identical BHPS variables in the Notes section of the variable view!

In this case, whilst the link to the instrument "GHQ" is retained in the new variable name, the variables are derived from information which is collected in the self-completion module of Understanding Society. These variables received the prefix "**sc**" in Understanding Society. Since the GHQ is a derived variable, it also received the "**_dv**" suffix. The questions underlying the GHQ summary measures on BHPS and Understanding Society have identical coding frames and collected in the same mode (self-completion) so they can safely be treated as the same. When merging the data, you will want to rename one of the set of variables, however.

A mentioned in the previous section, the naming conventions and construction of weights has changed from BHPS to Understanding Society. You may want to refer back to Worksheet 6 for a recap of weights in Understanding Society. If you compare the weights provided in the BHPS and in Understanding Society you will notice that Understanding Society provides more weights than the BHPS. For an overview of BHPS weights in Understanding Society, see the Understanding Society User Guide.

In this example we are planning to use data from adults aged 16+, including information collected in the self-completion questionnaire; our sample is the BHPS, and we are interested in analysing repeated cross-sections of data.

Question: Which weight do we use to match the weights we plan to use for the BHPS?

Theoretically, **w_indscbh_xw** is the most appropriate weight for this analysis because we are using data collected in the self-completion questionnaire module. In our example, however, the most appropriate weight will be **w_indinbh_xw**. There are no self-completion weights in BHPS; in order to be consistent across studies, we need to pick the weight with the lowest common denominator. The "lowest common denominator"-rule should also be applied with respect to other substantive variables.

A further consideration is that the BHPS provides weights separately for GB and the UK (from Wave 11 onwards) whereas Understanding Society only provides weights for UK. This is not a problem, however. You can use the UK weight and drop the Northern Ireland sample.

For the BHPS we have identified w**xrwght** (cross-sectional individual response weight, GB) and w**xrwtuk1** (cross-sectional individual response weight, UK) as appropriate weights for this exercise. Both are available in the INDRESP files.

Having established that BHPS and Understanding Society include identical markers of employment status and subjective well-being (we only need GHQ1) whilst also providing cross-sectional weights for individuals, we are now ready to load the BHPS data for Waves 1-18, do any data manipulations, and store them for merging with Understanding Society Wave 2 data.

9.3 CONSTRUCTING THE BHPS FILE AND PREPARING FOR MERGE WITH UNDERSTANDING SOCIETY

The Stata commands and data management strategies you have learned in Worksheets 1-6 equip you with all the skills necessary to generate the data set for this project. It is now just a case of applying the knowledge efficiently.

Start your do-file with the housekeeping commands and then define your working directories. In this project you will use data files from two different studies, which are contained in different working directories. You can easily refer to this other directory by creating an extra global macro containing the pathname of the directory

```
global dir2 "\\isernfs1\ConferenceData\bhps\"
```

Also create a new directory in your home area (M:\), which you call "example9data" where you can store data files you create for this project. You can do this interactively in Stata using the command **mkdir**. Alternatively, you can include this command in your do file. However, if you do that remember to add a capture command before this. Otherwise, when you run the do file second time round, it will give an error message saying that a folder with that name already exists.

```
capture mkdir M:\example9data
```

Create another global macro named "dir3" which refers to your new directory

```
global dir3 "M:\example9data\"
```

We now want to merge INDRESP data files from BHPS Wave 1-18. Seeing as there is a series of commands that are identical apart from the wave prefix, we can use the command **foreach** to load the data files, remove the wave prefix, generate the wave indicator, sort on the unique person time identifier and save the data files. All necessary commands and techniques have been introduced in previous examples.

Let's start by looking at the code for the last three waves of BHPS only.

```
foreach w in p q r  {
  use pid `w'hid `w'pno`w'hlghq1 `w'jbstat  ///
       `w'xrwght `w'xrwtuk1 using $dir2/`w'indresp, clear
       renpfix `w'
       capture rename id pid
       gen wave = strpos("abcdefghijklmnopqr","`w'")
       sort pid wave
       save $dir3/ind_junk`w', replace
       }
```

Note that the unique person identifier **pid** never has a wave prefix. This generates a problem for wave **p** where **`renpfix`** would rename **pid** to **id**. If you simply specified **`rename id pid`**, you would get an error message which would terminate the execution of the do-file, for all waves in which Stata would not find a variable called **id** (i.e. all waves other than wave p). Using **`capture`** tells Stata to swallow the error message and continue executing the do-file.

Now that you know the principle code, load the individual response records for each wave of the BHPS into working memory, remove the wave specific prefix, generate a wave-indicator called **wave**, and store the wave specific file in your folder for this project (i.e., example9data).

```
foreach w in a b c d e f g h i j k l m n o p q r {
  use pid `w'hidp `w'pno`w'hlghq1 `w'jbstat  ///
       `w'xrwght `w'xrwtuk1 using $dir2/`w'indresp, clear
       renpfix `w'
       capture rename id pid
       gen wave = strpos("abcdefghijklmnopqr","`w'")
       sort pid wave
       save $dir3/ind_junk`w', replace
       }
```

You will get the following error message

```
variable axrwtuk1 not found
r(111);
```

Stata aborts because it cannot find the variable **axrwtuk1** (i.e., the cross-sectional population weight for the UK) in data file AINDRESP; w**xrwtuk1** was only provided from Wave 11 onwards. You can loop over Waves 1-10 and Waves 11-18, separately.

```
foreach w in a b c d e f g h i j k {
  use pid `w'pno `w'hlghq1 `w'jbstat  ///
       `w'xrwght using $dir2/`w'indresp, clear
       renpfix `w'
       capture rename id pid
       gen wave = strpos("abcdefghijklmnopqr","`w'")
       sort pid wave
       save $dir3/ind_junk`w', replace
       }

foreach w in l m n o p q r {
  use pid `w'hlghq1 `w'jbstat  ///
       `w'xrwght `w'xrwtuk1 using $dir2/`w'indresp, clear
       renpfix `w'
       capture rename id pid
       gen wave = strpos("abcdefghijklmnopqr","`w'")
       sort pid wave
       save $dir3/ind_junk`w', replace
       }
```

Now use a similar **foreach** loop to append the data files from waves 1-17 to wave 18 data which is still in the memory

```
foreach w in a b c d e f g h i j k l m n o p q {
      append using $dir3/ind_junk`w'
      }
```

Last, but not least, label and save the data file.

```
lab dat "BHPS Waves1-18, long format"
save $dir3/ind_junk1to18, replace
```

You have now prepared the BHPS data set and can move on to preparing the Understanding Society data set.

9.4 LOAD UNDERSTANDING SOCIETY FILE AND PREPARE FOR MERGE WITH BHPS FILE

Load the Understanding Society Wave 2 data into the working memory.

```
use $dir/b_indresp, clear
```

Task: *Inspect this dataset. How many observations does it contain? How many variables? Which variable identifies the individuals and households? Can you find two variables that allow you to identify BHPS sample members?* Hint: Use **lookfor bhps sample.**

There are two ways to identify the BHPS respondents in the data file. All UKHLS data files contain the variables **pid** and **w_hhorig**. The variable **pid** is the unique person identifier which was assigned to BHPS sample members when they first joined the BHPS. The variable **w_hhorig** tells you the sample origin for all members of UKHLS, including four BHPS samples. Have a look at **b_hhorig**!

```
tab b_hhorig
```

The Wave 2 interim release contains information from six different samples, i.e, the (1) UKHLS general population sample in GB, (2) the UKHLS general population sample in Northern Ireland, (3) the BHPS GB sample, the BHPS regional boost samples for (4) Scotland and (5) Wales, and (6) the BHPS Northern Ireland sample. Wave 2 main release will furthermore include (7) the UKHLS Ethnic minority boost sample (GB only). Data for the Innovation Panel sample are released separately, see the dedicated section of the website http://www.understandingsociety.ac.uk/documentation/innovation-panel

You can use **w_hhorig** to keep only observations from the BHPS sample. Another way to select just BHPS sample members is by using the **pid**. To see how this might be done, inspect which values **pid** assumes for the new Understanding Society samples.

```
summ pid if (b_hhorig==1 | b_hhorig==2)
```

*Question: Which value does **pid** assume for new members of Understanding Society?*

The variable **pid** is inapplicable to members of new samples in Understanding Society. The codes for missing values are identical for BHPS and Understanding Society, i.e.,

-9 missing or wild
-8 inapplicable
-7 proxy or telephone interview
-2 refused
-1 don't know

Note that, technically, **pid** is an indicator fed forward from the last BHPS interview. Therefore, **pid** is not applicable also to new entrants to households of BHPS sample members. This is something that will change in the future to ease analysis of long runs of data. In particular, there are plans to re-release BHPS Waves 1-18; in the re-release every person who ever participated in BHPS will be issued with a **pidp** and linking files across studies will be easy.

In the meantime, generate a new variable (**pseudo_pid**= pidp*100) for new BHPS sample members.

```
gen double pseudo_pid=pidp*10 if pid==-8
```

Using the **double** qualifier of **generate** assures that Stata does not truncate the long numeric **pidp**. Then use **pseudo_pid** to replace inapplicable values for BHPS sample members.

```
replace pid= pseudo_pid if pid==-8 & (b_hhorig>=3 & b_hhorig<=6)
```

You could now either keep only observations with a valid **pid** or select on the sample origin indicator. We use the **b_hhorig** variable to keep only observations from the BHPS sample. Also drop all information that you are not interested in for this project. Make sure to keep the key linking variable **pid**! To link future waves of Understanding Society data you may want to retain that study's unique person identfier as well.

```
keep if b_hhorig>=3 & b_hhorig<=6
keep pid b_hidp b_pno pidp b_jbstat b_scghq1_dv b_indinbh_xw
```

Then remove the wave specific prefix and generate the wave identifier so that Understanding Society Wave 2 is treated as if it were BHPS Wave 19.

```
renpfix b_
gen wave = 17+strpos("abc","b")
lab wave "bhps sample wave"
sort pid wave
```

Inspect the data to see whether indicators are in the same format as on the BHPS.
Rename the GHQ variable so it has the same name as the BHPS variable. Do the same for any other variables you think are identical across the studies. You may want to undertake any other data cleaning operations at this stage as well, such as recoding missing values.

```
tab jbstat, miss
tab scghq1_dv, miss
rename scghq1_dv hlghq1
rename hidp hid

mvdecode _all, mv(-9/-1 97)
```

Were you struggling to generate your weighting variables so they refer to GB (**xrwght)** and UK (**xrwtuk1)**? **indinbh_xw** refers to the whole of the UK so you can just rename it to **xrwtuk1.** To create the GB population weight, copy **indinbh_xw** for the Great Britain samples only and name it **xrwght.**

```
gen xrwtuk1=indinbh_xw
lab var xrwtuk1 "Ind x-sctl person weight, UK"

gen xrwght=indinbh_xw if (hhorig>=3 & hhorig<=5)
lab var xrwght "Ind x-sctl person weight, GB"
```

You are now ready to label and store the data file in your project folder.

```
lab dat "BHPS Wave 19, long format"
save $dir3/ind_junk19, replace
```

9.5 MERGE BHPS WITH UNDERSTANDING SOCIETY FOR THE BHPS SAMPLE ONLY

With all data preparation done, this is very easy. Just append the Understanding Society Wave 2 data for the BHPS sample to the BHPS Wave1-18 data created in step (11.2).

```
use $dir3/ind_junk1to18, clear
append $dir3/ind_junk19
sort pid wave
```

Inspect the data to make sure the file looks like you intend.

*Question: Have you recoded **jbstat** in Wave 1? Better do it now! In future include this kind of data editing within the loop when you load in the data, i.e., before you append data files from individual waves as variable labels get overwritten with* **append**.

```
recode jbstat (5=6) (6=7) (7=8) (8=5) if wave==1
```

If that is all done, the data may be stored more efficiently, labelled and saved.

```
compress
lab dat "BHPS Waves 1-19, long format"
save $dir3/ind_1to19, replace
```

9.6 ANALYSIS

You can use the data created in this example, for instance, to look at mean GHQ for people living in GB 1991-2010, broken down by employment status. For 2001-2010 you can also look at the UK population as a whole. We will not discuss this here, but just to give you a flavour…

What is the mean GQH in the population GB and how has this changed over time?

```
bysort wave: table jbstat [pw= xrwght], content(mean hlghq1)
bysort wave: table jbstat [pw= xrwtuk1], content(mean hlghq1)
```

You can treat the data as pooled cross-sections, and predict GQH as a function of employment status

```
regress hlghq1 wave i.jbstat [pw= xrwght]
```

… or set the data up as panel data and estimate panel models such as random effects

```
xtset pid wave
xtreg hlghq1 wave i.jbstat
```

Note that we provide longitudinal weights for the BHPS sample in Wave 2 (2010) so it would be possible to undertake longitudinal analyses with correct weights. However, such analyses will only be possible for balanced samples, see **b_indin91_lw b_indin01_lw**.

9.7 FURTHER THINGS YOU MIGHT FIND USEFUL

BHPS interviews took place around September to December each year. In Understanding Society, interviews take place all year round. Always consider whether time of the year may be driving your results. E.g., whilst it may be true that children are, on average, unhappier with their school work in 2009 than they were in 1991-2008, this result could also be driven by the fact that some proportion of children in Understanding Society will be in the middle of their exams when they provide this information.

BHPS Wave 18 could also be treated as if it were Understanding Society Wave 1 for the BHPS sample within Understanding Society. Combining the weights is a tricky business, however. BHPS and Understanding Society refer to different, overlapping populations and we do not provide weights for this specific example. If you plan to undertake multivariate regression analysis and you do not care about representativeness, this is not an issue.

From Wave 3 onward, weights provided in the UKHLS data for the BHPS and UKHLS samples are combined and refer to the population that lived in the UK in 2009 (and results are representative for them in 2011/12, 2012/13 etc). Practically, when you use the BHPS sample only from this wave on, you would first define the survey features overall (i.e. use **`svyset`** including the weights ending on **ub_xw** or **ub_lw**) and then undertake analyses for just the subpopulation who are in the BHPS samples. See **`help svyset`**, and refer to Example 6 for using the svy-suite of commands in Stata. Hint: You would follow a similar approach if you wanted to analyse sample members from two overlapping waves in a particular calendar year.

Whilst there are some members of minority ethnic groups in the BHPS sample of Understanding Society, they do not form part of the ethnicity strand of Understanding Society.

The way in which the relationship of household members to each other is collected, changed in the UKHLS compared to BHPS. The variable **w_relationship** on data file EGOALT tells you the relationship of ego to alter; w**rel** on the BHPS tells you alter's relationship to ego.