An Introduction to the WERS-REPONSE Stata dataset

Version 1.0 (May 2016)

1. Introduction

The WERS-REPONSE Stata dataset ('the WR dataset' hereafter) was compiled as part of a research project to comparatively analyse workplace employment relations in Britain and France. The comparative analysis was based on data from the British *Workplace Employment Relations Surveys* (WERS 2004 and 2011) and the French *Enquête Relations Professionnelles et Négociations d'Entreprises* (REPONSE 2005 and 2011). The project was funded by the Leverhulme Trust under Research Project Grant RPG-2013-399 and the research findings were published as:

> Amossé T, Bryson A, Forth J and Petit J (eds.) (2016) *Comparative Workplace Employment Relations: An Analysis of Britain and France*, Basingstoke: Palgrave Macmillan.

A major aim of the project was to create an accessible research infrastructure that would facilitate and encourage further comparative analysis beyond that which is presented in the aforementioned book. To that end, the project website[1] provides access to a comprehensive set of resources that will enable researchers who have obtained the original survey datasets both to replicate and to extend the analysis that is presented in the book.

Those resources include five syntax files (Stata do files) which, when run as a group, combine cases from WERS and REPONSE (MQ and SEQ, 2004/5 and 2011) into a single dataset, and derive many of the 'harmonised' variables that were used in the comparative analysis. The purpose of this document is to provide a 'user guide' to those syntax files and the resulting data file.

The project website also provides access to translated versions of the WERS and REPONSE 2011 questionnaires (in PDF format), and to a spreadsheet (in Microsoft Excel format) which provides a detailed listing of all topics covered in the respective Management and Employee Surveys from the 2004/2011 WERS and the 2005/2011 REPONSE. The spreadsheet also details the names of relevant variables in each of the source datasets.

Users can use the syntax files and associated resources to re-compile the WR dataset used by Amossé *et al* when writing the aforementioned book. The full WR dataset can then be fully utilised (and potentially extended) in the user's own comparative analysis. Alternatively, users can treat the syntax and associated resources as a reference tool to support their own data compilation efforts. In this case, the Excel spreadsheet and translated questionnaires can be used to identify topics of interest, and the Stata syntax files can be consulted to see how those source variables might be combined to create single, harmonised data items for the analysis of specific topics of interest.
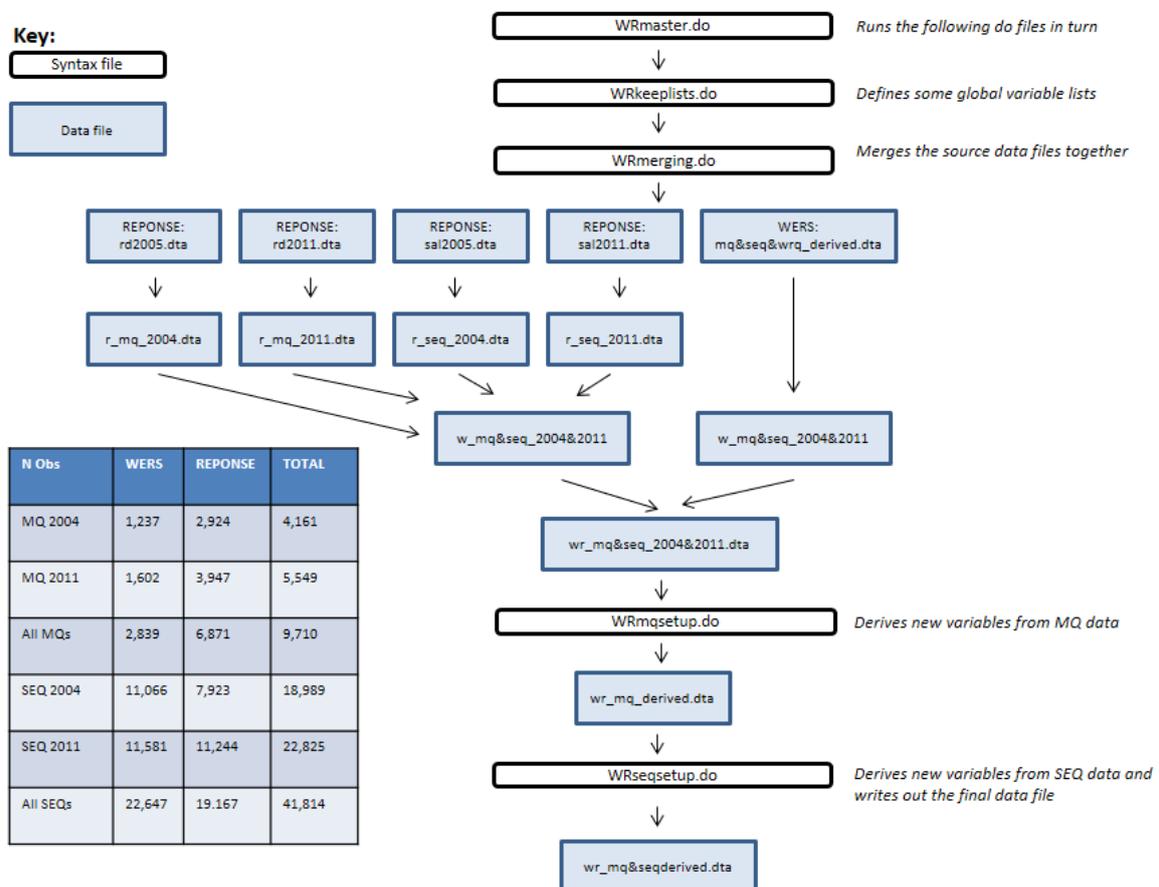
Users who with to replicate specific results – or to extend specific pieces of analysis – presented in Amossé *et al* should contact the research team if they are unsure how the WR dataset has been employed analytically in the writing of the book. Contact details are provided at the end of this user guide.

---

[1] URL: http://www.niesr.ac.uk/projects/employment-relations-britain-and-france

2. Overview of the data compilation process

An overview of the data compilation process is provided in Figure 1 below. In brief, the syntax takes the source data (MQ and SEQ only) from WERS 2004 and 2011, and the equivalent data from REPONSE 2005 and 2011, and merges them into one single data file. It then derives variables from the MQ data for each survey, and follows this by deriving variables from the SEQ data for each survey. Finally, the syntax writes out a new data file (`wr_mq&seqderived.dta`) which comprises all of these components (source data and derived variables, for the MQ and SEQ elements of WERS 2004 and 2011, and the MQ and SEQ elements of REPONSE 2005 and 2011). Indicator variables and variable naming conventions are used in such a way as to allow the user to easily navigate these various components of the combined dataset.[2]

Figure 1: WERS&REPONSE Data Production Process

Key:
- Syntax file
- Data file

WRmaster.do	*Runs the following do files in turn*
WRkeeplists.do	*Defines some global variable lists*
WRmerging.do	*Merges the source data files together*

REPONSE: rd2005.dta → r_mq_2004.dta
REPONSE: rd2011.dta → r_mq_2011.dta
REPONSE: sal2005.dta → r_seq_2004.dta
REPONSE: sal2011.dta → r_seq_2011.dta
WERS: mq&seq&wrq_derived.dta

w_mq&seq_2004&2011
w_mq&seq_2004&2011

wr_mq&seq_2004&2011.dta

| WRmqsetup.do | *Derives new variables from MQ data* |

wr_mq_derived.dta

| WRseqsetup.do | *Derives new variables from SEQ data and writes out the final data file* |

wr_mq&seqderived.dta

N Obs	WERS	REPONSE	TOTAL
MQ 2004	1,237	2,924	4,161
MQ 2011	1,602	3,947	5,549
All MQs	2,839	6,871	9,710
SEQ 2004	11,066	7,923	18,989
SEQ 2011	11,581	11,244	22,825
All SEQs	22,647	19.167	41,814

3. What is the source (input) data?

The input files are:

- For WERS: A combined WERS 2004 and 2011 data file containing all of the WERS data deposited with the UK Data Service (`mq&seq&wrq_derived.dta`). This data file can be created by employing the syntax that has been deposited at the UK Data Service by the

---

[2] Users should note that the final dataset is approximately 285Mb in size.

WERS 2011 Research Team (the syntax will be sent out along with the data to anyone who successfully applies for access to WERS 2011 via the UK Data Service).[3] That WERS syntax is similar in structure to the WR syntax described here: it takes the source data from WERS 2004 and WERS 2011 and creates a single combined data file which additionally contains some derived variables.

- For REPONSE: The individual source data files from REPONSE 2005 and 2011 (`rd2005.dta`, `rd2011.dta`, `sal2005.dta` and `sal2011.dta`). These data files have not been deposited in a public data archive, but can be obtained from the French Ministry of Labour.[4]

The WR syntax assumes that all of these data files are stored somewhere on your local computer. The relevant file paths can be specified in `WRmaster.do`.

4. Which source variables are contained within the dataset?

The dataset does not attempt to include all of the source variables from the WERS and REPONSE MQ and SEQ datasets. Instead, the dataset only includes those source variables that were required for the comparative analysis undertaken by the research team.

The dataset does not contain any cases or variables from the WRQ in either WERS or REPONSE. The WRQ data were judged by the research to be insufficiently comparable and were not used in the comparative analysis.

The full set of source variables that are included in the WR data file is specified in the syntax file `WRkeeplists.do`.

5. Which survey observations does the dataset contain?

The dataset contains all of the cases that fall within the coverage of the harmonised WERS-REPONSE population, that is:

- In 2004, the MQ sample contains all trading sector workplaces with 21+ employees operating in NACE Rev 1.1 Sections D (Manufacturing) to O (Other Services), with the exception of Section L (Public Administration). The 'trading sector' is defined as all private sector workplaces plus those of trading public-sector corporations.
- In 2011 the MQ sample contains all trading sector workplaces with 11+ employees operating in NACE Rev 2 Sections C (Manufacturing) to S (Other Services), with the exception of Section O (Public Administration).
- In 2004 and 2011, the REPONSE SEQ sample includes only those employees who were recorded on the centralised DADS administrative register as having 15+ months of tenure at the establishment.[5] The SEQ sample for WERS includes those employees who answered the tenure question (A1) and reported that they had 12+ months of tenure.

---

[3] The UKDS catalogue record for WERS 2011 can be found here:
https://discover.ukdataservice.ac.uk/catalogue/?sn=7226&type=Data%20catalogue
[4] The REPONSE pages on the website of the French Ministry of Labour can be found here: URL: http://travail-emploi.gouv.fr/etudes-recherches-statistiques-de,76/statistiques,78/relations-professionnelles,85/les-enquetes-relations,280/l-enquete-reponse-2010-2011,17939.html
It should be noted, however, that access to the REPONSE data is subject to approval from the Conseil National de l'Information Statistique (URL: www.cnis.fr).
[5] DADS: Déclaration Annuelle de Données Sociales.

The dataset does not contain any other cases (e.g. public sector workplaces in WERS; smaller workplaces in WERS; employees with <12 months of tenure in WERS; employees that were surveyed in REPONSE but whose establishments did not provide an MQ interview; or observations from surveys undertaken prior to 2004).

These inclusion/exclusion criteria are applied to the data in `WRmerging.do`.

6. **Variable naming conventions**

To avoid any conflicts over variable names when compiling the consolidated data file, a set of naming conventions has been applied in the WR syntax:

- Any source variables from WERS are given the prefix "w_" (e.g. `w_nestwt`)
- Any source variables from REPONSE are given the prefix "r_" (e.g. `r_pds_etab`)
- Any new, harmonised variables derived from MQ data are given the prefix "x" (e.g. `xestwt`)
- Any new, harmonised variables derived from SEQ data are given the prefix "y" (e.g. `yseqwt`)

Note: many of the source variables in REPONSE have names written in capital letters. These are all converted to lower-case in `WRmerging.do`.

7. **What is the structure of the final WR dataset?**

Figure 2 below gives an overview of the data structure within the output data file (`wr_mq&seqderived.dta`). The key point to note is that the dataset contains distinct records for each observation (each eligible workplace and each eligible employee) from WERS and REPONSE (2004 and 2011, MQ and SEQ). Each workplace in the dataset has a unique serial number (`serno`). Each employee is uniquely identified through the combination of `serno` and `persid`.

Figure 2: Structure of the WR dataset

			serno	persid	r_siren_neweff	...	w_zallemps	...	xsurvey	xyear	xmq	xseq	xallemps	...
WERS	2004 MQ	2004 SEQ	1	1			10		1	0	1	1	10	
		2004 SEQ	1	2			10		1	0	0	1	10	
			..	..	..		..		..	..	..	..	..	..
	2011 MQ	2011 SEQ	51	1			13		1	0	1	1	13	
		2011 SEQ	51	2			13		1	0	0	1	13	
			..	..	..		..		..	..	..	..	..	..
REPONSE	2004 MQ	2004 SEQ	101	1	34		.		0	1	1	1	34	
		2004 SEQ	101	2	34		.		0	1	0	1	34	
			..	..	..		..		..	..	..	..	..	..
	2011 MQ	2011 SEQ	151	1	16		.		0	1	1	1	16	
		2004 SEQ	151	2	16		.		0	1	0	1	16	
			..	..	..		..		..	..	..	..	..	..

Notes:
(i) The order of variables on the WR datafile does not correspond precisely to this stylised illustration
(ii) The serial numbers (serno, persid) depicted in the diagram are also illustrative

In order to manage this combined dataset in analysis, a series of indicator variables are provided to identify different types of cases. These are listed in Table 1 below.

Table 1: Indicator variables

Subset	Indicator variable
All observations from WERS	`xsurvey==1`
All observations from REPONSE	`xsurvey==0`
All observations from 2004	`xyear==0`
All observations from 2011	`xyear==1`
All MQ observations	`xmq==1`
All SEQ observations	`xseq==1`
All observations from workplaces with 21+ employees (a consistent employment threshold for 2004 and 2011)	`x0411==1`
All MQ and SEQ observations from workplaces that featured in the 2004-11 panel sample	`xpanel==1`
All panel sample observations from workplaces with 21+ employees in both 2004 and 2011	`xp0411==1`
All panel sample observations from workplaces with 21+ employees in 2004 and 2011 which also generated SEQ observations in both years	`xps0411==1`

Having compiled the WR dataset, the following Stata code would thus provide you with an (unweighted) mean of workplace size among those workplaces observed in 2011, separately for WERS and REPONSE.

```
mean xallemps if xyear==1 & xmq==1, over(xsurvey)
```

Similarly, the following Stata code would thus provide you with the (unweighted) percentage of employees that are male in each year of the REPONSE data, among the consistently-defined subset of workplaces with 21+ employees.

```
mean ymale if xsurvey==0 & xseq==1 & x0411==1, over(xyear)
```

Table 2 provides an overview of the numbers of observations in the WR dataset as an aid for users.

Table 2: Numbers of observations in the WR dataset (cross-section)

Full cross-sectional samples: **2004/2011 samples with consistent employment threshold:**

21+ emps in 2004, 11+ emps in 2011   21+ emps in 2004, 21+ emps in 2011

All MQ obs (xmq==1):

	WERS	REPONSE	Total
2004	1,237	2,924	4,161
2011	1,602	3,947	5,549
Total	2,839	6,871	9,710

All MQ obs (x0411==1 & xmq==1):

	WERS	REPONSE	Total
2004	1,237	2,924	4,161
2011	1,289	3,321	4,610
Total	2,526	6,245	8,771

All SEQ obs (xseq==1):

	WERS	REPONSE	Total
2004	11,066	7,923	18,989
2011	11,581	11,244	22,825
Total	22,647	19,167	41,814

All SEQ obs (x0411==1 & xseq==1):

	WERS	REPONSE	Total
2004	11,066	7,923	18,989
2011	10,288	10,009	20,297
Total	21,354	17,932	39,286

All obs in dataset:

	WERS	REPONSE	Total
2004	11,407	8,175	19,582
2011	12,073	11,566	23,639
Total	23,480	19,741	43,221

All obs (X0411==1):

	WERS	REPONSE	Total
2004	11,407	8,175	19,582
2011	10,680	10,222	20,902
Total	22,087	18,397	40,484

Table 2: Numbers of observations in the WR dataset (panel)

2004/2011 samples with consistent employment threshold:

2004/2011 samples with consistent employment threshold and SEQs in both years:

21+ emps in 2004, 21+ emps in 2011

21+ emps in 2004, 21+ emps in 2011, and SEQs in both years

All MQ obs (xp0411==1 & xmq==1):

	WERS	REPONSE	Total
2004	447	840	1,287
2011	447	840	1,287
Total	894	1,680	2,574

All MQ obs (xps0411==1 & xmq==1):

	WERS	REPONSE	Total
2004	256	749	1,005
2011	256	749	1,005
	512	1,498	2,010

All SEQ obs (xp0411==1 & xseq==1):

	WERS	REPONSE	Total
2004	4,352	2,363	6,715
2011	3,986	2,648	6,634
Total	8,338	5,011	13,349

All SEQ obs (xps0411==1 & xseq==1):

	WERS	REPONSE	Total
2004	3,245	2,292	5,537
2011	3,226	2,501	5,727
	6,471	4,793	11,264

All Panel obs in dataset (xp0411==1):

	WERS	REPONSE	Total
2004	4,455	2,428	6,883
2011	4,111	2,683	6,794
Total	8,566	5,111	13,677

All Panel obs in dataset (xps0411==1):

	WERS	REPONSE	Total
2004	3,245	2,292	5,537
2011	3,226	2,501	5,727
	6,471	4,793	11,264

Note: the full WR dataset also contains those panel cases with 21+ emps in 2004 and 11+ emps in 2011 (xpanel==1)
There are 26 such workplaces in REPONSE and also 26 in WERS

8. Weights

Table 2 below lists the names of the weighting variables on the combined dataset. The weighting variables are compiled from those weights that have been made available with the source data. For instance, in the case of any MQ observations that originate from REPONSE 2011, the new variable 'xestwt' utilises the weighting values provided in the variable 'pds_etab' in the source data file 'rd2011.dta'.

The only adjustment made to these weighting values in the WR syntax is to add scaling factors, which are necessary to ensure that each survey-year makes an equal contribution to any pooled analysis.

Table 2: Names given to the main weighting variables

Name	Description	Use
xestwt	MQ establishment weight (2004 or 2011)	Analysis of MQ data for 2004 or 2011 with a workplace-base (showing the % of workplaces with a particular characteristic)
xempwt	MQ employment weight (2004 or 2011)	Analysis of MQ data for 2004 or 2011 with an employment-base (showing the % of employees in workplaces with a particular characteristic)
yseqwt	SEQ weight (2004 or 2011)	Analysis of SEQ data for 2004 or 2011 (showing the % of employees with a particular characteristic)
xestwt0411	MQ establishment weight (2004 and 2011, wps with 21+ emps)	Analysis of MQ data (with workplace-base) for 2004 and 2011 among workplaces with 21+ employees
xempwt0411	MQ employment weight (2004 and 2011, wps with 21+ emps)	Analysis of MQ data (with employment-base) for 2004 and 2011 among workplaces with 21+ employees
yseqwt0411	SEQ weight (2004 and 2011, wps with 21+ emps)	Analysis of SEQ data for 2004 and 2011 among workplaces with 21+ employees
xpqestwt	MQ panel establishment weight (2004 and 2011)	Analysis of MQ panel data for 2004 and 2011 (workplace-base)
xpqempwt	MQ panel employment weight (2004 and 2011)	Analysis of MQ panel data for 2004 and 2011 (employment-base)

Having compiled the WR dataset, the following Stata code would thus provide you with a *weighted* mean of workplace size among workplaces observed in 2011, separately for WERS and REPONSE.

```
        mean xallemps if xyear==1 & xmq==1 [pw=xestwt], over(xsurvey)
```

Similarly, the following Stata code would provide you with the *weighted* percentage of employees that are male in each year of REPONSE, among the consistently-defined subset of workplaces with 21+ employees.

```
        mean ymale if xsurvey==0 & xseq==1 & x0411==1 [pw=yseqwt0411],
        over(xyear)
```

9. Labels, missing values and other coding conventions

The source data for WERS and for REPONSE 2005 are typically well-labelled. However, the SAS source files that were available to us for REPONSE 2011 contained few variable labels or value labels. We have not attempted to label these REPONSE source variables. However, we have attempted to label all of the new derived variables that we have created.

There are some variations in the way in which missing values are coded across the file. The WERS source data have separate missing values for 'Refused' (.a), 'Don't know' (.b), 'Not applicable' (.e) and so on. The REPONSE source data has Refused/Don't know as valid values (the code for which can change from question to question); all filtered cases (where a person is skipped around a question because it is not applicable) are coded as 'system missing'.

When creating new derived variables, we have tried to assign a valid code or a user-missing value to each observation. In many cases, we will have followed the WERS coding system by default, such that a workplace/employee would be coded ".a" on the derived variable if they were asked the question but did not give a response, and coded ".e" if they were filtered around the question.

New variables which have only two valid values (say "yes" and "no"; or "high" and "low") are typically coded in binary or 'dummy variable' form (0,1).

10. First steps

A user wishing to compile the WR dataset should proceed as follows:

- Contact the French Ministry of Labour and request access to the REPONSE data for 2005 and 2011 (note: it may take some weeks before access is granted). Once the data files have been supplied, convert them from SAS to Stata using Stat/Transfer or an equivalent program.
- Register with the UK Data Service (https://www.ukdataservice.ac.uk/) and download the WERS 2004 and 2011 general-use data files (SN 5294 and 7226 in the UKDS data catalogue). The data should be downloaded in Stata format.
- Compile the WERS 2004+2011 combined dataset using the Stata syntax provided by the WERS Research Team with the 2011 WERS deposit.
- Edit 'WRmaster.do' so that the file paths suit your local setup.
- Run 'WRmaster.do' to create the WR data file. The production process will run automatically as illustrated in Figure 1 and the resulting data file (wr_mq&seqderived.dta) will be written out to the folder identified on the 'path_derived' global in 'WRmaster.do'.

11. The contributors

The Leverhulme research team comprised eight researchers:

- Thomas Amossé
- Philippe Askenazy
- Alex Bryson
- Christine Erhel
- John Forth
- Héloïse Petit
- Antoine Rebérioux
- Zinaida Salibekyan.

Any queries regarding this syntax – or about the broader project from which it originates – may be directed to John Forth (j.forth@neisr.ac.uk) or Thomas Amossé (Thomas.amosse@cee-recherche.fr).