

**GUIDE TO THE ANALYSIS OF  
THE WORKPLACE EMPLOYMENT  
RELATIONS SURVEY 2004**

[Version 1.5: 18<sup>th</sup> June 2008]

John Forth, Simon Kirby & Lucy Stokes

**WERS 2004 Information and Advice Service**

 National Institute Of Economic And Social Research

2 Dean Trench Street, Smith Square, London SW1P 3HE

Tel: +44(0) 20 7654 1933 E-mail: [wers2004@niesr.ac.uk](mailto:wers2004@niesr.ac.uk)  
URL: <http://www.wers2004.info>

# Contents

<b>INTRODUCTION .....</b>	<b>4</b>
<b>1. OBTAINING DOCUMENTATION AND DATA.....</b>	<b>5</b>
1.1 How do I obtain documentation about WERS 2004, such as copies of the survey questionnaires?.....	5
1.2 How do I obtain the WERS 2004 data to conduct my own analyses?.....	5
1.3 Are there any data items from WERS 2004 that have not been deposited at the UK Data Archive?.....	5
1.4 What are the benefits of using the data held at the ONS Virtual Micro-data Laboratory?.....	6
1.5 Can I interrogate the WERS 2004 data without using a sophisticated statistical software package?.....	7
1.6 Where can I find more information about the additions and revisions to the WERS data made in April 2007?.....	7
1.7 Have any additional data items been linked to the WERS 2004 data?.....	8
1.8 How was the panel data file revised in January 2008?.....	11
1.9 What's new in the WERS Time-Series dataset deposited in April 2008?.....	12
<b>2. FINDING YOUR WAY AROUND THE DATA FILES .....</b>	<b>13</b>
2.1 What are the variable naming conventions in WERS 2004?.....	13
2.2 Do the records in the WERS 2004 data files have unique identifiers?.....	15
2.3 How do the WERS 2004 data files record responses from questions that permitted multiple responses?.....	15
2.4 How are missing values coded in WERS 2004?.....	16
2.5 Are there any problems that I should be aware of in the data files?.....	17
2.6 What are the differences between the 1998 and 2004 survey questionnaires?.....	17
<b>3. MANIPULATING THE DATA FILES.....</b>	<b>19</b>
3.1 How do I combine the separate Cross-Section data files together for linked analysis?.....	19
3.2 How do I construct a panel dataset including data from 1998 and 2004?.....	20
<b>4. DERIVED VARIABLES.....</b>	<b>23</b>
4.1 Are you making any derived variables available, to avoid duplication of effort between researchers?.....	23
4.2 Do you have ready-made syntax that combines the T values with the categorical response in questions about workforce proportions (e.g. BINVMANG and BINVMANT)?.....	23
4.3 Why isn't there a derived variable to combine COFFJOB and COFFJOB?.....	24
4.4 How do I create a dummy variable from a multiple response set?.....	24
<b>5. WEIGHTING AND STATISTICAL INFERENCE IN WERS 2004 .....</b>	<b>25</b>
5.1 What are the names of the weighting variables in the WERS data files?.....	25
5.2 Why should I weight my analyses?.....	27
5.3 How does weighting work?.....	28
5.4 Why will standard procedures give me incorrect standard errors?.....	29
5.5 What software packages allow me to properly account for the WERS sample design (both through weighting and correct estimation of standard errors)?.....	30
5.6 How do I apply weights and correctly estimate variances in Stata?.....	31
5.7 How do I apply weights and correctly estimate variances in SPSS?.....	35
5.8 Does it matter whether one uses Stata's svy commands or SPSS Complex Samples module?.....	39
5.9 What adjustments should I make if I am analysing only a subset of the full dataset or my variables have missing values?.....	39
5.10 Colleagues tell me that a fully-specified regression model doesn't need weighting – are they right?.....	41
5.11 Why are all of the weights in WERS scaled to sum to 100 and how does this affect my analysis?.....	41

5.12 Why should I not use <i>iweights</i> (in Stata) or <i>WEIGHT BY</i> (in SPSS)? .....	42
5.13 Where can I find more information about weighting and statistical inference? .....	43
5.14 Where can I find more information about the revised weight for the Cross-Section Survey of Employees? .....	43
<b>6. PUBLISHING YOUR RESEARCH .....</b>	<b>45</b>
6.1 How do I acknowledge the use of the WERS 2004 data in publications? .....	45
6.2 How can I make others aware of my research outputs? .....	46

## *Introduction*

### **Introduction**

This *Guide to the Analysis of WERS 2004* brings together the advice and guidance contained within the Frequently Asked Questions section of the WERS 2004 Information and Advice Service web-site at: <http://www.wers2004.info/FAQ.php>

The FAQs contain numerous web links which remain active in this pdf version.

The FAQs are updated on a periodic basis as new questions arise from users. On each occasion, this Guide will also be updated to include the new material. The publication of new FAQs, and subsequent revisions of this Guide, will be advertised via our mailing list.

[Sign up to the WERS mailing list](#)

## **1. Obtaining documentation and data**

### ***1.1 How do I obtain documentation about WERS 2004, such as copies of the survey questionnaires?***

Electronic versions of the survey questionnaires and other source documents for the survey (such as the Technical Report) are provided on the WERS 2004 Information and Advice Service website, in the section describing the nature and conduct of WERS 2004. [Visit these pages.](#)

We have also compiled an eight-page introduction to WERS 2004 (updated in April 2007), which is available to download here as an Acrobat pdf file: [Introduction to WERS 2004.](#)

### ***1.2 How do I obtain the WERS 2004 data to conduct my own analyses?***

The publicly available data files from WERS 2004 are available from the UK Data Archive at the University of Essex. For further details, visit our webpage on [how to obtain the data files.](#)

### ***1.3 Are there any data items from WERS 2004 that have not been deposited at the UK Data Archive?***

Non-deposited data can be categorised into two groups: permanently deleted data items and restricted data items.

The permanently-deleted data items contain information that would directly identify a workplace or respondent, such as the name of the workplace or the organization to which it belonged. An example would be **AORGNAM**E in the Cross-Section Survey of Managers. These data items have been permanently deleted in order to preserve the anonymity of respondents. The assurances made to respondents prior to their participation in the survey mean that these data items can never be made publicly available.

Prior to April 2007, some additional data items were restricted, in line with assurances to respondents that access would be withheld until two years following the final

## *1. Obtaining documentation and data*

interview. This included region identifiers and detailed industry codes, which are considered to present a further risk to respondent anonymity, and numerical data on workplace performance, which is considered to be particularly sensitive. These items have been added to the data available from the [Economic and Social Data Service](#) from April 2007.

Note that in workplaces where managers consented to the linking of WERS data with other data sources (**MLINKDAT=Yes**), the Office for National Statistics has linked data from WERS 2004 to the Annual Business Inquiry, the Annual Survey of Hours and Earnings and the Business Structure Database. These linked data will only ever be available via the [Virtual Microdata Laboratory](#) operated by the Business Data Linking section at the Office for National Statistics. They will not be made available via the Economic and Social Data Service, due to ONS restrictions on access to the data. Further details on these and other linked data are provided in [FAQ 1.7](#).

### ***1.4 What are the benefits of using the data held at the ONS Virtual Micro-data Laboratory?***

As noted in [FAQ 1.3](#), in workplaces where managers consented to the linking of WERS data with other data sources (**MLINKDAT=Yes**), the Office for National Statistics linked the Cross-Section and Panel data from WERS 2004 with longitudinal data on organisation performance collected as part of the Annual Business Inquiry. These linked data will only ever be available via the [Virtual Microdata Laboratory](#) operated by the Business Data Linking section at the Office for National Statistics.

Users wishing to access the linked WERS-ABI data can therefore only do so at the Virtual Microdata Laboratory. While data from the Cross-Section Financial Performance Questionnaire (FPQ) has been available from the Economic and Social Data Service from April 2007, using the linked WERS-ABI data can allow users to fill gaps in the FPQ data, therefore increasing the number of observations available for analysis. Another advantage is that the ABI sometimes provides more timely data. For further information please see [WIAS Technical Paper 1: Objective Data on Workplace Performance](#).

Data from WERS 2004 have also been linked to the Annual Survey of Hours and Earnings (ASHE) and separately to the ONS Business Structure Database (BSD). The linked WERS-ASHE data and the linked WERS-BSD data are also only available at the Virtual Microdata Laboratory. Further details of these linked datasets are provided in the answer to [FAQ 1.7](#).

Unbanded versions of the Travel to Work Area variables (see [FAQ 1.7](#)) have also been made available at the ONS Virtual Microdata Laboratory, along with the postcodes of each individual surveyed workplace.

***1.5 Can I interrogate the WERS 2004 data without using a sophisticated statistical software package?***

Simple frequency tables are available on the WIAS web-site for each question in the Cross-Section Survey of Managers, the Cross-Section Survey of Employees and the Survey of Employee Representatives. [Visit tabulations](#).

The data from WERS 2004 are also available via Nesstar, the on-line analysis tool provided by the Economic and Social Data Service. [Visit WERS 2004 pages on Nesstar](#). This allows access to the data via a simple web-based interface, providing a number of useful functions, such as the ability to compile cross-tabulations, view frequencies, produce graphs and conduct correlation and regression analyses. The data from WERS 1998 are also available via Nesstar. [Visit WERS 1998 pages on Nesstar](#).

Those interested in the findings from WERS 2004 may also wish to consult the extensive primary analysis of the survey data, published in three separate reports totalling over 500 pages. [Details of primary analysis](#).

***1.6 Where can I find more information about the additions and revisions to the WERS data made in April 2007?***

The original WERS 2004 data files held by [ESDS](#) were updated in April 2007 to incorporate additional data, along with a number of small revisions.

The newly available data include regional and industry identifiers for both the Cross-Section Management data file and the Panel data file, along with the data file relating

## *1. Obtaining documentation and data*

to the Cross-Section Financial Performance Questionnaire. These data items were withheld from general release until April 2007 in order to maintain the anonymity of respondents, following assurances that were made to respondents during fieldwork.

At the same time as depositing this additional data, the opportunity was taken to make some small revisions to the WERS 2004 datasets. These include a small [revision to the weight for the Cross-Section Survey of Employees](#), along with some small additional revisions.

Full details of both the additions and revisions to the data are contained within the introductory note accompanying the data in the [ESDS on-line catalogue](#), which is also available to [download](#) from the WIAS website.

### ***1.7 Have any additional data items been linked to the WERS 2004 data?***

#### ***Data on financial performance:***

The Office for National Statistics has linked the Cross-Section and Panel data from WERS 2004 with longitudinal data on organisation performance collected as part of the Annual Business Inquiry. These linked data are available via the [Virtual Microdata Laboratory \(VML\)](#) operated by the Business Data Linking section at the Office for National Statistics.

Two [technical papers](#) describe the linked data and compare the objective values with the subjective ratings collected in the Cross-Section Management Interview.

The Office for National Statistics has also made available links from the Inter-Departmental Business Register (IDBR) to company reference numbers (CRN) used within the 2004 [FAME database](#). These CRNs are available for around 1,300 establishments in the 2004 Workplace Employment Relations Survey (WERS), providing users with the mechanism to link WERS to FAME data for these workplaces. It is noted that the link between CRNs and WERS workplaces are made at the level of the enterprise through enterprise reference numbers. Therefore, workplaces in WERS that belong to the same enterprise will each be linked to the same CRN. A note providing further details of this linkage has been prepared by the ONS Virtual Microdata Laboratory (VML); this is available to download [here](#).

Prospective users of the VML should note that FAME is a commercial dataset and a general-use version of the FAME dataset is not available at the VML. Instead, users wishing to analyze linked data from WERS and FAME would need to bring their own FAME dataset to the VML, where it could be linked to the WERS data, using the link variable provided.

***Data on the local labour market:***

In order to facilitate research which takes into account characteristics of the local area in which the establishment is located, and after consulting with users, we have made available a small number of variables at Travel to Work Area (TTWA) level. These data were made available from the [Economic and Social Data Service](#) in October 2007.

The data include unemployment and vacancy rates, along with the ratio of unemployment to vacancies, for matching on to the WERS 2004 Cross-Section Survey of Managers and the 1998-2004 Panel datasets. Similar data are also available for matching on to the WERS98 Cross-Section Management Questionnaire, following users' request for consistent TTWA variables for 1998 and 2004. These replace the earlier data file local98.\*, previously available from the Economic and Social Data Service. It was not possible, retrospectively, to identify the derivation of variables in local98.\*, or to replicate their values.

An urban-rural indicator, and a variable indicating the percentage of the TTWA population belonging to an ethnic minority group, are also available for matching on to the WERS 2004 Cross-Section Management Questionnaire.

We have sought to provide the data at as disaggregated level as possible. However, it has been necessary to band the variables in some cases, in order to prevent the deduction of TTWA identities via unique values. Unbanded versions of the variables have been deposited at the ONS [Virtual Microdata Laboratory](#). To facilitate further geographical matching, the postcodes of each individual surveyed workplace have also been made available at the VML.

Further information is provided in the documentation accompanying the data in the [ESDS online catalogue](#), and is also available to download below.

[WERS 2004 Travel to Work Area variables: note for users](#)

## *1. Obtaining documentation and data*

[WERS 1998 Travel to Work Area variables: note for users](#)

[WERS 1998 -2004 Panel, Travel to Work Area Variables: note for users](#)

### ***Data on wages:***

The Annual Survey of Hours and Earnings (ASHE) is an annual survey in which employers report on the working hours and earnings of around 165,000 randomly-selected employees. The survey provides detailed information about components of wages (basic wages, overtime pay, incentive payments, pensions), which is linkable over time. The ONS have identified around 6,000 ASHE records originating from workplaces that participated in the WERS 2004 Cross-Section Survey. WERS then provides a great deal of contextual information about the employee's workplace that can be used to help understand variations in earnings within the ASHE data. The linked dataset is available for analysts within the ONS Virtual Microdata Laboratory (see [FAQ 1.4](#)). The characteristics of the matched dataset are described in [WIAS Technical Paper 3: Linking the Annual Survey of Hours and Earnings to WERS 2004](#).

Further information on ASHE is provided on the [ONS website](#)

### ***Data on business structures:***

The ONS Business Structure Database (BSD) is a version of the Inter-Departmental Business Register that is prepared for research use. It provides information on an enterprise's legal status and country of ownership, among other things. It also provides information on demographic events including birth, death, merger and takeover. Many of these data are available at establishment level. The ONS have successfully matched the WERS establishment records to the BSD, providing BSD information for 2,143 (93%) of the workplaces in the WERS 2004 Cross-Section. The linked dataset is available for analysts within the ONS Virtual Microdata Laboratory (see [FAQ 1.4](#)).

### ***Data on occupations:***

A series of variables containing information on the composition of occupations (at the SOC (2000) Unit Group level) for employees in Great Britain were made available

from the Economic and Social Data Service in January 2008. These variables have been deposited with the purpose of providing contextual information about the occupations of employees surveyed within the WERS 2004 Survey of Employees, based on information provided in the Labour Force Survey and other sources about the average characteristics of employees in those occupations. The occupational data can be matched to the WERS 2004 Survey of Employees, and include the following variables:

Gender composition

Proportion under the age of 22

Proportion aged 50 or more

Ethnic composition

Proportion with a long-term work-limiting disability

Average gross hourly wages

Proportion with highest educational qualification at each NVQ-equivalent level

National Statistics Socio-Economic Classification

Female and male CAMSIS scores

Further details are provided in the documentation accompanying the data in the [ESDS online catalogue](#), and are also available to download [here](#).

### ***Data at industry-level***

The [WERS time-series dataset](#) contains a link variable (TEUKLEMS) that enables observations to be linked to the EU KLEMS industry-level database, which provides measures of economic growth, productivity, employment creation, capital formation and technological change at the industry level for the UK from 1970 onwards. Further information about EU KLEMS can be found at the [EU KLEMS Project web page](#) .

### ***1.8 How was the panel data file revised in January 2008?***

The 2004 panel data file has been revised to include previously absent employment data for 88 cases. These data were made available from the Economic and Social Data Service in January 2008.

## *1. Obtaining documentation and data*

Further details are provided in the documentation accompanying the data in the [ESDS online catalogue](#), and are also available to download [here](#).

### ***1.9 What's new in the WERS Time-Series dataset deposited in April 2008?***

The WERS Time-Series Dataset is formed from the interviews with the main management respondent in each of the five cross-section surveys (1980, 1984, 1990, 1998 and 2004).

Aside from the addition of a 2004 time-point, the latest version of the dataset (Version 2.3) is largely the same as the original version (v1.1) and the subsequent minor revision (v1.2) deposited with the UK Data Archive in May 2002. However, a small number of improvements have been made to enhance the consistency of variable definitions and to expand the range of variables incorporated within the dataset.

Further information about the Time-Series dataset can be found [here](#).

## 2. Finding your way around the data files

### 2. Finding your way around the data files

#### 2.1 What are the variable naming conventions in WERS 2004?

In general, each variable name has two parts: a one or two-character prefix that signifies which section of the relevant questionnaire the variable arises from, while the remaining characters are intended to give some sense of the topic covered by the question. Variables arising from questions that permitted multiple responses have a number at the end to signify the order of response.

##### *Variables in xs04\_mq.\*:*

A one-character prefix signifies the section of the Main Management questionnaire from which the variable arises. So **ASINGLE** arises from Section A of the questionnaire. Variables arising from multiple response questions are numbered from 1 upwards (or, from 01 if 10 or more responses were permitted), so that **AHOWCH01** contains the first numeric response given by a particular manager to the question about changes of ownership, and **AHOWCH12** the twelfth response. Note, however, that few respondents gave the maximum number of responses to any multiple response question; in most cases they mentioned only one or two items from the code list. In these instances, the remaining variables in the set will be empty (i.e. 'system missing').

##### *Variables in xs04\_erq.\*:*

Variables arising from the Employee Representative questionnaire have a two-character prefix. The first character (W) is short-hand for Worker Representative. The second character signifies the section of the questionnaire from which the variable arises. So **WAREPTY** arises from Section A of the Employee Representative questionnaire. Variables arising from multiple-response questions are labelled in the same way as in **xs04\_mq.\***

##### *Variables in xs04\_seq.\*:*

A one-character prefix points to the relevant section of the Survey of Employees questionnaire. Questions inviting more than one box to be ticked (**E4**, **E7**, **E8** and **E9**) yield one dichotomous variable for each of the possible responses (e.g. **E4\_1** to

## 2. Finding your way around the data files

**E4\_6**). An additional variable with the same name as the question (**E4** in this example) indicates the number of boxes ticked by the respondent. For questions that have a number of elements to them (e.g. **A6**), and in which the respondent is invited to provide an answer for each element or row, letters (a, b etc.) are appended to the variable name to identify the response to each row (e.g. **A6a** to **A6d**).

Note: Some questions in the SEQ were not intended to elicit multiple responses but were multi-coded by a number of respondents: this applies to **B2**, **D2**, **E10** and **E14**. In this case, the main variable takes the value of -6 if more than one box was ticked. Additionally, one dichotomous variable for each of the possible responses is included in the file, for example, **B2MULT1** to **B2MULT8**, plus an additional variable **B2MULT** which indicates the number of boxes ticked by the respondent.

### *Variables in xs04\_fpq with derived variables.\*:*

Details of variable names in the Financial Performance Questionnaire data file are provided in the introductory note accompanying the dataset. The variable names for the raw data contained in the deposited data file are largely descriptive. All derived and edited variables are pre-fixed with the letter 'N'; while variables beginning with 'xcode' are overcodes.

### *Variables in ps9804\_pq04.\*:*

Similarly to the Cross-Section Management data file, variables in the panel data file have a one-character prefix that signifies the section of the Panel Survey Management Questionnaire from which the variables arises. So **ASINGLE** arises from Section A of the questionnaire. Variables arising from multiple response questions are also labelled in the same way as in **xs04\_mq.\***

Note that most of the variable names in the 2004 Panel Survey data file are identical to those appearing in the WERS 1998 Cross-Section Survey of Managers, which constitutes the first wave of the panel. This makes the combination of the two waves of data into a single data file problematic. However, syntax which has been compiled for this purpose is outlined in [FAQ 3.2](#).

### ***2.2 Do the records in the WERS 2004 data files have unique identifiers?***

Each workplace in the Cross-Section Survey has a unique identifier (**SERNO**). This can be used to link responses from managers, employee representatives and employees at the same establishment. For example, one can combine information from **xs04\_mq.\*** with that from **xs04\_erq.\*** in order to compare managers' and employee representatives' reports of the climate of employment relations at the workplace. Alternatively, one might combine information from **xs04\_mq.\*** with that from **xs04\_seq.\*** in order to assess the degree to which employees' attitudes vary by industry or size of workplace. The process of matching of data from different data files using SPSS or Stata is outlined in [FAQ 3.1](#).

In addition, within the Cross-Section, each employee representative is uniquely identified by the combination of **SERNO** and **WAREPTYP**. Each employee is uniquely identified by the combination of **SERNO** and **PERSID**.

Each workplace in Wave 2 of the Panel Survey has a unique identifier, also called **SERNO**, which can be used to link the data to responses in Wave 1 of the Panel - the 1998 Cross-Section Survey (deposited separately at the UK Data Archive under study number 3955). [FAQ 3.2](#) gives instructions on how to compile a single data file containing data from both waves of the Panel Survey.

### ***2.3 How do the WERS 2004 data files record responses from questions that permitted multiple responses?***

Some questions in the Cross-Section Survey of Managers, the Cross-Section Survey of Employee Representatives, and the Panel Survey of Managers, permitted multiple responses (e.g. **CFACTORS** in the Cross-Section Survey of Managers). These questions are identified with the symbol ^ in the questionnaires. In the data files, the responses to these questions are stored in successive variables (e.g. **CFACTOR1**, **CFACTOR2** and so on). The first variable in the set contains the first response mentioned in the interview, the second variable holds the second response, etc.

Some multiple-response questions allowed the respondent to give answers other than those included on the pre-specified code frame. In such cases, the 'other' answers were recorded verbatim by interviewers and then subsequently coded into additional

## *2. Finding your way around the data files*

variables beginning with the letter X (e.g. **XCFACT1**, **XCFACT2**, **XCFACT3**). These additional variables follow the main set on the data file. The original ‘other’ code remains present in the main set of variables. A respondent using the ‘Other, please specify’ item on **CFACTORS** (code 9) will thus have a value of 9 in one of the variables **CFACTOR1-CFACTOR9** plus a set of values in **XCFACT1-XCFACT3** which correspond to the codes assigned to their verbatim answers during the data coding and editing process. The original ‘other’ code should be ignored during analysis to prevent double-counting.

In the Cross-Section Survey of Employees, a small number of questions in Section E of the questionnaire were designed to accommodate multiple-responses (i.e. **E4**, **E7**, **E8**, **E9**). In these cases, the primary variable indicates the number of responses given by each respondent at that question, and subsequent variables indicate whether each of the available response categories was chosen. For example, the variable **E4** indicates the number of responses given by each respondent at question **E4**, whilst codes indicating whether each of the six possible were chosen are contained in **E4\_1** (“No dependent children”) to **E4\_6** (“Children aged 12-18”).

In addition, a number of questions in the Survey of Employees attracted multiple responses from small numbers of employees, despite the questionnaire indicating that the respondent should tick only one box. In these cases, the main variable is coded to a specific missing value (-6). For questions that attracted multiple responses from more than 20 employees (**B2**, **D2**, **E10** and **E14**), the multiple responses have been made available on the dataset adjacent to the main variable.

### ***2.4 How are missing values coded in WERS 2004?***

The standard approach to the coding of missing values in WERS 2004 is to use the same three codes in each data file: -9 (Not answered / Refused); -8 (Don’t know); and -1 (Not applicable, i.e. Not asked). A small number of variables have additional missing values to identify particular circumstances: an example is the use of -6 on single-coded questions in the SEQ that were inadvertently multi-coded by respondents (e.g. **B2**).

## 2. Finding your way around the data files

In the SPSS data files, these codes are designated as ‘user missing values’ and so are automatically excluded from any statistical procedures, such as the production of means. However, in the Stata data files, the codes are not designated as missing values and they will therefore be treated as valid values by any statistical procedures. The syntax command cited below can be used to designate the three standard codes (-9, -8 and -1) as ‘missing values’ in the Stata data files. Such codes will then be excluded by statistical procedures, such as `-tabulate-` or `-summarize-`. However, as Stata stores missing values as infinitely large positive values, the clause “`if var<.`” will still be required if these cases are to be ignored by transformation commands such as `-generate-`.

```
mvdecode _all, mv(-9=.a \-8=.b \-1 = .c)
```

Note that the treatment of missing values was slightly different in WERS 1998. In those data files, cases that were not asked a particular question were left as ‘system missing’. ‘Not answered / Refused’ was coded 8 (or 98 in variables with more than seven valid codes, 998 in variables with more than 97 valid codes and so on). ‘Don’t know’ was coded 9 (or 99 etc).

### ***2.5 Are there any problems that I should be aware of in the data files?***

We have produced a set of ‘variable notes’ that alert users to specific issues or problems with particular variables that ought to be borne in mind when conducting analysis of the survey data. These include such things as routing errors, defective text-fills and instances of ambiguous question wording that have become apparent during primary analysis of the survey data. We encourage users to report further issues to us by emailing [wers2004@niesr.ac.uk](mailto:wers2004@niesr.ac.uk). Updates to these variable notes will be advertised via our mailing list.

[Access variable notes.](#)

### ***2.6 What are the differences between the 1998 and 2004 survey questionnaires?***

Users of WERS 1998 may be seeking to use WERS 2004 for similar avenues of research. Some of this may be longitudinal in nature, where the research focuses on

## *2. Finding your way around the data files*

comparability of data items. Some research will use innovations and improvements in WERS 2004, for example, taking advantage of improved question wording or new data items. To aid these research activities, we have documented the substantive changes to the Cross-Section Management Questionnaire and the Survey of Employees Questionnaire, providing a reference guide to users who are already familiar with WERS 1998, or seeking to conduct comparative analysis.

Both sets of equivalence tables can be downloaded by clicking on the links below.

[Equivalence tables for MQ](#)

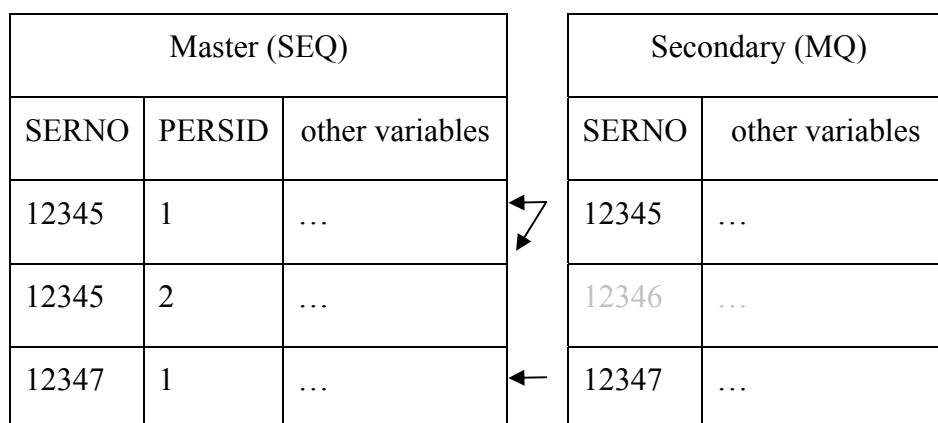
[Equivalence tables for SEQ](#).

### 3. Manipulating the data files

#### 3.1 How do I combine the separate Cross-Section data files together for linked analysis?

There are a number of different reasons why one might want to link files from the Cross-Section together. For instance, users may wish to add data from the management data file into the Employee Representative data file, in order to compare the responses from managers and employees representatives in the same workplaces. Alternatively, one might wish to add data from the Management data file into the Survey of Employees data file, in order to allow workplace data to be used in the analysis of employee responses. These are both ‘one-to-many’ matches, and either of these tasks can be achieved using the `MATCH FILES` command in SPSS or the `-merge-` command in Stata. The two files are combined by matching on the unique workplace identifier (**SERNO**) which is present on both data files.

The diagram below indicates the nature of the matching process, with data from a secondary file (in this case the MQ) being matched onto data from a master file (in this case the SEQ).



When conducting such a match, one is unlikely to want the ‘non-matched’ record for workplace 12346 to appear in the resultant matched file (i.e. one wishes to add only variables to the master file, not cases). To ensure that 12346 is ignored, SPSS users should specify the secondary file as a ‘look-up table’ on the `MATCH FILES` command, whilst Stata users should follow the `-merge-` command with a further command to: `drop _merge==2.`

### *3B3. Manipulating the data files*

A further possibility is that one may wish to link the employee data file with the Management or Employee Representative data in order to use some summary measure of employee attitudes or behaviour in analyses conducted at workplace level. For example, one may wish to summarise employees' perception of management's attitude towards trade unions for use in the analysis of union recognition. The process is essentially the same but, before the `MATCH FILES` or `-merge-` commands can be used, a summary measure must first be produced for each workplace from the employee data. This is because these commands do not perform 'many-to-one' matches, whereby multiple employee observations are matched onto a single management or employee representative record. Summary measures at workplace level may be produced from the employee data using the `AGGREGATE` command in SPSS or the `-collapse-` command in Stata.

### ***3.2 How do I construct a panel dataset including data from 1998 and 2004?***

The data that comprise the 1998-2004 Panel Survey are deposited in two separate data files which need to be linked prior to analysis.

The first wave of data comes from the 1998 Cross-Section Survey and consists of observations on 2,191 workplaces with 10 or more employees; this data file has study number 3955 in the UK Data Archive's on-line catalogue. The second wave, conducted as part of WERS 2004, consists of a small number of data items for each of these 2,191 workplaces (essentially indicating its survival status and employment level in 2004) and a larger set of data items arising from repeat interviews at 938 of the surviving workplaces. These data are contained within the Panel Survey data file that can be found in the single WERS 2004 deposit at the UK Data Archive (study number 5294). [Obtaining data.](#)

The data from 1998 and 2004 may be combined to form a single data file with either of two possible structures:

1. **One record per case (wide form).** In this format, each workplace occupies a single record (row) in the resulting data file. One block of variables contains data from 1998 and another block contains data from 2004. The two sets of variables might be distinguished by a different prefix (so **XASINGLE** might

contain the data arising from the question named **ASINGLE** in 1998, whilst **YASINGLE** might contain the data from **ASINGLE** in 2004).

2. **One record per year (long form)**. In this format, each workplace has two records in the data file. The first record contains data from 1998 and the second contains data from the same workplace in 2004. An index variable indicates the year to which each row relates. The number of variables is reduced by at least half in comparison with the first structure because: (a) the 1998 and 2004 values for consistently-defined data items can be contained within a single variable; and (b) questions or variables that are not consistent across the two years of the panel do not feature in the resulting data file.

A 1998-2004 panel dataset in wide form can be constructed from the deposited data files using either the `MATCH FILES` command in SPSS or the `-merge-` command in Stata. A panel dataset in long form can be constructed using the `VARSTOCASES` command in SPSS, or the `-reshape-` command in Stata.

We have compiled syntax to perform both operations in either SPSS or Stata. This can be downloaded [here](#). The syntax requires SPSS v12 or Stata v7, or later versions of either software, as it creates some variables with names that exceed 8 characters. The steps that are followed in either package are essentially the same:

To construct a panel dataset in ‘wide form’:

1. Open the 1998 data file and fix known errors using the syntax compiled by the WERS98 Data Dissemination Service ([available here](#)). Then append the variables contained in the 2004 data file. As part of the matching process, variables containing 1998 data are prefixed with the letter X and those containing 2004 data are prefixed with the letter Y. Note that many variables will now have names that exceed 8 characters (permitted only in SPSS version 12.0 or later, and Stata version 7 or later).
2. Enforce consistency between the 1998 and 2004 blocks of variables in respect of the codes used for missing values: -1 for “Not asked” and -9 for “Don’t know or Not answered”. Note that the distinction between “Don’t know” and “Not answered” is sacrificed to simplify the assimilation process. A small

### 3B3. Manipulating the data files

number of variables with idiosyncratic missing value codes (e.g. **XXCODES**, **YXCODES**, **YEMPS2004**) are left untouched. [This step is omitted in Stata as there is no simple recode available]

3. Enforce consistency between the 1998 and 2004 blocks of variables in the names of otherwise equivalent variables. The practice is to make the 2004 data items consistent with 1998.
4. Save the new data file.

To construct a panel dataset in ‘long form’:

1. Follow the steps required to construct a ‘wide’ panel dataset.
2. Drop any workplaces that were not re-interviewed in 2004 (since those workplaces that were not interviewed have very limited information in 2004 and thus almost no consistent data items across the two years).
3. Drop any variables that are not entirely consistently across the two years (users may add further variables, after manipulating them to ensure consistency).
4. Remodel the data file to create a second row for each workplace, and populate this row with data from the 2004 block of consistent variables.
5. Create an index variable to identify the first row for each workplace as containing data from 1998 and the second as containing data from 2004.

Note: The panel compilation syntax calls upon Mq98fix.sps / Mq98fix.do. These files can be downloaded from the web-site of the [WERS98 Data Dissemination Service](#).

## 4. Derived variables

### ***4.1 Are you making any derived variables available, to avoid duplication of effort between researchers?***

Yes, we are. The first batch of variables is described under [FAQ 4.2](#) below.

The second batch of derived variables are those used in the compilation of the [First Findings booklet](#). These files contain the syntax developed by the WERS Research Team in the preparation of First Findings; much of the syntax also provides the starting point for the analysis presented in the sourcebook. The files include the syntax needed to replicate all of the figures reported in First Findings, as well as the syntax to create the many derived variables used in the analysis. For the 2004 Cross-Section Survey of Managers alone, the syntax creates more than 200 derived variables. Note that all derived variables are provided as SPSS syntax files.

Thirdly, a number of derived variables have been added to the data file for the Financial Performance Questionnaire (FPQ). These derivations are set out in an SPSS syntax file, WERS\_2004\_FPQ\_derived\_variables.sps. This syntax file is contained within the zip file containing the WERS 2004 data and documentation, deposited with the [Economic and Social Data Service](#)

### ***4.2 Do you have ready-made syntax that combines the T values with the categorical response in questions about workforce proportions (e.g. BINVMANG and BINVMANT)?***

In the Cross-Section Survey of Managers, questions that asked about the proportion of employees to which a practice applied (e.g. **BINVMANG**) could be answered in two ways. The first was for the respondent to cite the percentage using the code frame provided (i.e. 'All (100%)', 'Almost all (80-99%)' etc). The second was for the respondent to cite the actual number of employees, with this value being coded into a second variable with a T suffix (in this case **BINVMANT**). SPSS syntax which combines the responses by recoding the T values back into the percentage code frame is provided below.

**Syntax to combine T variables and categorical responses:**

[WERS 2004 XS MQ - derived variables v1.sps](#)

#### **4.3 Why isn't there a derived variable to combine COFFJOB and COFFJOB?**

**COFFJOB** collects the number of experienced employees in the largest non-managerial occupation (LOG) who have received off-the-job training. However, in order to code this response back to the categories used in **COFFJOB**, one requires data on the total number of experienced employees in the LOG to provide a denominator for the fraction. This data was not collected in the questionnaire. In hindsight, the code 97 should not have been included as an option on **COFFJOB**.

#### **4.4 How do I create a dummy variable from a multiple response set?**

The example given here takes the multiple response set **CFACTORS** from the Cross-Section Survey of Managers (comprising variables **CFACTOR1-CFACTOR9** and **XCFAC1-XCFAC3**) and computes a (1,0) dummy variable named **CFAC\_AGE** to identify those workplaces in which the management respondent chose code 5 'Age'.

To create the dummy variable in SPSS, type:

```
do if missing(cfactor1)=0 .  
+ compute cfac_age=any(5,cfactor1 to cfactor9,xfact1 to xfact3) .  
end if .
```

To create the dummy variable in Stata, type:

```
egen cfac_age=eqany(cfactor1-cfactor9 xfact1-xfact3), values(5)  
recode cfac_age 0=. if cfactor1<1
```

The resulting dummy variable has 293 cases coded 1 and 2,000 cases coded 0. Those two respondents that did not answer **CFACTORS** remain missing on the dummy.

## 5. Weighting and statistical inference in WERS 2004

In contrast to many surveys of individuals, business surveys such as WERS are commonly based upon sample designs that involve many significant departures from the principle of simple random sampling which underpins most standard statistical procedures. Adjustments must be made to those procedures if the analyst is to produce unbiased results and make correct inferences about the statistical precision of their estimates. Below we provide answers to a number of FAQs about weighting and statistical inference.

### 5.1 What are the names of the weighting variables in the WERS data files?

The weights for the 2004 Cross-Section Survey are as follows:

**ESTWTNR** – the standard establishment weight, which should be used to produce workplace-level estimates from the Survey of Managers (e.g. the proportion of all workplaces that recognise trade unions).

**EMPWTNR** – an alternative weight for the Survey of Managers that can be used to produce analyses that indicate the proportion of all employees to whom a particular workplace characteristic applies (e.g. the proportion of all employees working in establishments with recognised trade unions). **EMPWTNR** is simply the product of **ESTWTNR** and **ZALLEMPS** (the total number of employees working at the establishment).

**FPQWTNR** – the standard establishment weight for the subset of cases returning a Financial Performance Questionnaire. Based on **ESTWTNR** with an adjustment for differential non-response.

**SEQWTNR** – the standard weight for the Survey of Employees, which accounts for variations in sampling fractions at both stages of the design (i.e. selection of the sample of establishments, and selection of the sample of employees within each establishment). **Users should note that a small revision to this weight was made in April 2007.** Further details are contained in the answer to [FAQ 5.14](#).

**WRQWTNR** – the standard weight for the Survey of Employee Representatives. Users should note that **WRQWTNR** is simply a re-scaled

## 5. Weighting and statistical inference in WERS 2004

version of ESTWTNR. The reason for re-scaling the WRQ weight was so that the weighted base summed to 100 when analysing the full WRQ sample.

[Further details](#).

The weights for the 1998-2004 Panel Survey are as follows:

**ESTWTNR** – establishment weight, which should be used to produce workplace-level estimates of the 2004 outcome for the full sample of 2,191 establishments from 1998.

**PQWTNR** – establishment weight, which should be used to produce workplace-level estimates for the sub-sample of 938 establishments that were interviewed in 2004. This weight should be used to produce estimates from either Wave 1 (1998) or Wave 2 (2004) among this sub-sample.

All weights are scaled so that the sum of weighted cases is 100. Further information on the sample design and the methods used to construct the weights are provided in Section 7 of the [WERS 2004 Technical Report](#).

### *Revised weights for WERS 1998:*

Improvements to the methods used to construct the weights for WERS 2004, in comparison to those used in WERS 1998, mean that direct comparisons between the original 1998 Cross-Section (UKDA study number: 3955) and those establishments with 10 or more employees in the 2004 Cross-Section may be impaired by differences in the weighting strategies. However, **ESTWTNR** on the Panel Survey data file has been produced for all 2,191 cases from WERS 1998 using the improved method. This weight can be matched onto the original data from WERS 1998 and used in place of **EST\_WT** to produce Cross-Section estimates for 1998 that are weighted on a comparable basis to estimates derived from the 2004 Cross-Section. All of the 1998 estimates presented in the WERS 2004 First Findings and the full sourcebook *Inside the Workplace* have been produced using **ESTWTNR** rather than **EST\_WT**.

**ESTWTNR** can be used directly to weight data from the 1998 Survey of Managers and Survey of Worker Representatives. A weight for the 1998 Survey of Employees can be derived by multiplying the original SEQ weight variable (**EMPWT\_NR**) by a

factor that is equal to the ratio between the original and revised establishment weights. A further scaling factor equal to 100/1084 is used to scale the sum of the weights to 100. The syntax is as follows:

SPSS:

```
compute seqwtnr=((empwt_nr*(estwtnr/est_wt))*100/1084).
```

Stata:

```
gen seqwtnr=((empwt_nr*(estwtnr/est_wt))*100/1084).
```

### ***5.2 Why should I weight my analyses?***

In short, the design of the WERS 2004 sample has the effect of introducing bias to any estimates that are derived from the raw data. As a result, one must account for the sample design by applying weights to the data, if one wishes to obtain unbiased population estimates.

When drawing a simple random sample, each member of the population has the same probability of selection and so, in the absence of non-response biases, the achieved sample will inevitably resemble the population from which it was drawn. However, when the sample of workplaces was drawn for the WERS Cross-Section, large workplaces (which are relatively uncommon in the population) were purposefully given a higher probability of selection than smaller workplaces and workplaces from less populated industries (such as Electricity, Gas and Water Supply) were over-sampled relative to those from more heavily populated industries (such as Wholesale and retail). The consequence was that the profile of the issued sample of workplaces was out of kilter with the population at large, since large workplaces and those from small industries were proportionately over-represented.

Similarly in the employee survey, once an employee's workplace had been selected to participate in WERS, a member of staff in a small workplace had a higher probability of receiving a Survey of Employees Questionnaire than an employee in a large workplace (since questionnaires were distributed to all employees in workplaces with 5-25 employees and to only 25 employees in larger workplaces). So employees from

## 5. *Weighting and statistical inference in WERS 2004*

small workplaces were over-represented in the employee sample when compared with the population for the employee survey (i.e. all employees in workplaces participating in the WERS Cross-Section).

On top of biases introduced purposefully as part of sampling, variable rates of non-response can cause the achieved sample to depart in additional ways from the population it is intended to represent. In the workplace survey, smaller workplaces had a lower response rate on average than larger workplaces. In the employee survey, men were less likely to respond than women.

Weights equal to  $1/(\text{probability of selection \& response})$  are used during analysis to bring the profiles of the achieved samples of workplaces and employees into line with the profiles of the respective populations, thereby removing known biases introduced by the sample selection and response process. Most software packages – including Stata and SPSS - allow for the use of such weights in analysis.

For more information on the WERS sample design see Section 2 of the [WERS 2004 Technical Report](#). The derivation of weights is described in Section 7 of the Technical Report.

The theory of weighting is described more fully here:

<http://www.dcs.napier.ac.uk/peas/theoryweighting.htm>

### **5.3 *How does weighting work?***

If we take the workplace sample as an example, each workplace is assigned a weight value that reflects its degree of over or under-representation in the achieved sample when compared with the population at large. If one assigned a weight of 1 to a workplace that is neither under nor over-represented, then workplaces that are under-represented in the sample would have weights greater than 1 and those that are over-represented would have weights less than 1. A weighted estimator (e.g. to compute a mean) uses these weight values in combination with the actual values of the variable being analysed to obtain unbiased estimates. Whereas a standard estimator gives each case equal weight in the computation of the mean value, a weighted estimator will give more importance to cases with larger weights.

Most software packages – including Stata and SPSS - allow for the use of weights in analysis.

The mechanics of weighting are described more fully here:

<http://www.dcs.napier.ac.uk/peas/theoryweighting.htm>

#### ***5.4 Why will standard procedures give me incorrect standard errors?***

The standard textbook formulae for estimating the variance (and thus the standard error) of a statistic (whether it be a mean, a proportion or a regression coefficient) assume that the statistic has been derived from a simple random sample drawn with replacement (SRSWR). Under SRSWR, the sample is drawn by taking a random selection of cases from the population using a constant sampling fraction. Each case is also made available for re-selection at every draw, even if it has already been sampled (hence the term ‘with replacement’).

WERS 2004 incorporated a number of departures from the SRSWR design. Standard methods of estimating standard errors are therefore no longer valid. The departures and their general impact on standard errors are specified below.

In the workplace survey:

Stratification of the population prior to sampling (tends to give smaller standard errors than SRSWR)

Unequal sampling fractions across strata (tends to give larger standard errors than SRSWR)

Sampling without replacement (tends to give smaller standard errors than SRSWR)

Post-stratification (tends to give smaller standard errors than SRSWR)

Additionally in the survey of employees:

Clustering such that employees were only selected from workplaces participating in the workplace survey (tends to give larger standard errors than SRSWR)

Whilst some of these departures tend to give smaller standard errors than a SRSWR design, these effects are outweighed by the impact of using unequal sampling fractions and clustering, so that the overall effect for most data items in the survey is

## 5. Weighting and statistical inference in WERS 2004

to increase standard errors when compared with a SRSWR design. A statistic called the ‘design factor’ (DEFT) gives a measure of the degree of amplification in sampling errors that results from using a complex sample design rather than SRSWR. NatCen calculated a median DEFT of 1.45 among a range of estimates from the WERS Management Questionnaire, indicating that standard formulae will underestimate the size of standard errors from the WERS MQ by approximately 45% on average. The median DEFT for the Survey of Employees Questionnaire was 1.59, giving a general indication of the additional impact of clustering.

The use of standard formulae for variance estimation will therefore usually lead us to conclude that estimates from the WERS 2004 are more precise than they really are (so-called Type II errors). Hence, one needs to use alternative methods of estimating standard errors which account for the more complex sample design used in WERS. These alternatives include ‘linearization’ and ‘replication methods’. A small number of software packages include such methods of variance estimation, including Stata and SPSS.

Further details on the WERS 2004 sample design are provided in Section 2 of the [WERS 2004 Technical Report](#). For more on DEFTS, see Section 8 of the Technical Report.

For a fuller discussion of design effects and the correct estimation of standard errors in complex surveys, see: <http://www.napier.ac.uk/depts/fhls/peas/errors.asp>

### ***5.5 What software packages allow me to properly account for the WERS sample design (both through weighting and correct estimation of standard errors)?***

Stata and SPSS are the most commonly used software packages among WERS analysts. Both allow for the application of weights and for the correct estimation of standard errors. However, both employ special procedures to do so.

In **Stata**, these procedures have the prefix `svy` and are described in the dedicated Stata 9 Reference Manual on Survey Data. The `svy` procedures are part of the standard software package from Version 5.0 onwards.

In **SPSS**, the procedures are contained within a separate module called ‘Complex Samples’ which is available as an add-on to the standard Base module from Version 12.0 onwards.

Note: Both Stata and SPSS also allow users to apply weights when using statistical procedures that do not simultaneously produce correct variance estimates. The specification of `iweights` in Stata, and the use of the `WEIGHT BY` command in SPSS, will weight the data appropriately, ensuring unbiased coefficients. However, the standard errors arising from any statistical commands issued under these conditions will invariably be incorrect. Only use of the `svy` suite of commands in Stata or the Complex Samples module in SPSS will ensure that coefficients and standard errors are both correctly computed. See also [FAQ 5.12](#).

Stata and SPSS are not the only software packages to correctly account for complex sample designs. A summary of available software for the analysis of complex surveys is provided here: <http://www.hcp.med.harvard.edu/statistics/survey-soft/>

### ***5.6 How do I apply weights and correctly estimate variances in Stata?***

Since Version 5.0, Stata has included a special suite of procedures for the analysis of data from complex samples. These procedures have the prefix `svy` and are described fully in the dedicated Stata 9 Reference Manual on Survey Data.

The analyst first uses the `svyset` command to inform Stata about the survey weights and the various features of the sample design. Use of the `svy:` prefix before estimation commands such as `mean`, `proportion` or `logit` then ensures that Stata takes account of these features of the sample design during estimation to produce weighted estimates and correct standard errors. In Stata 9, the `svy:` prefix replaces the dedicated `svy` commands that were present in earlier versions of the software (`svymean`, `svyprop`, `svylogit` etc).

The various features of the WERS sample design are specified on the `svyset` command in the following way:

## 5. Weighting and statistical inference in WERS 2004

### ***When analysing the WERS 2004 MQ:***

```
svyset [pweight=estwtnr], strata(wpstr04) vce(linearized)
```

#### ***where:***

estwtnr is the workplace weight variable

wpstr04 identifies the stratum from which the workplace was sampled

### ***When analysing the WERS 2004 ERQ:***

```
svyset [pweight=wrqwtnr], strata(wpstr04) vce(linearized)
```

#### ***where:***

wrqwtnr is the employee representative weight variable

wpstr04 identifies the stratum from which the workplace was sampled

### ***When analysing the WERS 2004 FPQ:***

```
svyset [pweight=fpqwtnr], strata(wpstr04) vce(linearized)
```

#### ***where:***

fpqwtnr is the workplace weight variable for the subset of cases returning  
Financial Performance Questionnaires

### ***When analysing the WERS 2004 SEQ:***

```
svyset serno [pweight=seqwtnr], strata(wpstr04)  
vce(linearized)
```

#### ***where:***

seqwtnr is the employee weight variable

serno is the unique workplace identifier (designating the clustering of the  
employee survey sample)

### ***When analysing the WERS 1998-2004 Panel Survey:***

```
svyset [pweight=pqwtnr], strata(wpstr98) vce(linearized)
```

**where:**

`pwgwtnr` is the weight variable for continuing workplaces

`wpstr98` identifies the stratum from which the workplace was sampled in 1998

***Where do I find these commands in Stata's menu system?***

The `svyset` command is accessed via:

Statistics | Survey data analysis | Setup & utilities | Declare survey design for dataset

The various estimation commands are available via:

Statistics | Survey data analysis

***Where do I get `wpstr04` and `wpstr98`?***

Syntax to derive these additional variables is provided in the table at the end of this FAQ. These variables will be added to subsequent versions of the data files distributed by the UKDA.

***Why don't the `svyset` commands specify a finite population correction (`fpc`)?***

It is possible to add this term, but it has little practical effect on standard errors. In our trials, standard errors deviated at only the 3rd decimal place when specifying a finite population correction for the MQ dataset (computed as the number of achieved observations divided by the population total, with the latter indicated in Table 2.1 in the [WERS 2004 Technical Report](#)). Use of the term is also questionable when using non-random subsets of the full data set. It might be noted that NatCen (the fieldwork company for WERS) did not use this feature when computing design effects for the WERS Technical Report.

***Why don't the `svyset` commands above take account of post-stratification?***

It is possible to split the final weight variable (e.g. `ESTWTNR`) into two separate components: that part which accounts for the use of unequal sampling fractions and that part which post-stratifies the achieved sample to match the profile of the population (see Section 7 of the [WERS 2004 Technical Report](#) for a discussion of how the weights in WERS are derived). It should then be possible to specify these two

## 5. Weighting and statistical inference in WERS 2004

components separately on the `svyset` command, using the `pweight` and `postweight()` options. However, this is only feasible when estimating on the full sample, since population totals for specific sub-samples are not known. Additionally, in our trials we have found it difficult to get Stata to weight even the full sample properly under this arrangement. In practice, the reduction in standard errors that is theoretically achievable by separately specifying the post-stratification element of the sample design is relatively small: in our trials it altered standard errors at only the 3rd decimal place. It might be noted that NatCen (the fieldwork company for WERS) did not use this feature when computing design effects for the [WERS 2004 Technical Report](#) (see p.106).

### ***Linearization vs jackknife:***

Stata offers two alternative algorithms for estimating variances: linearization and the jackknife replication method. Interested readers are referred to Chapter 9 of Sharon Lohr's book (see [FAQ 5.13](#)) for an exposition of the differences. Suffice it to say that, in our tests, the linearization method proved to be much quicker and standard errors estimated under the two alternative methods usually varied at only the 3rd or 4th decimal place. We would therefore recommend the use of linearization over the jackknife.

### ***Obtaining correlations or population standard deviations:***

Stata does not contain an `svy` command to produce bivariate correlations. However, the procedure for obtaining correlations which take account of the survey design is outlined on the Stata web-site: [obtaining correlations](#).

Similarly, there is no direct means of obtaining estimates of the standard deviation of a variable within the wider population. But again, a procedure for obtaining this statistic is outlined on the Stata web-site: [obtaining population standard deviations](#).

### ***Sample output:***

The sample output provided in the table below includes a worked example of the application of the `svyset` procedure to run a logistic regression using the WERS 2004 MQ data. The output also indicates the impact of specifying only certain parts of

the sample design. And it indicates how the `svy` output differs from that obtained using `iweights` or unweighted data.

**Stata do file to create wpstr04:**

[WERS 2004 Cross-Section - wpstr04.do](#)

**Stata do file to create wpstr98:**

[WERS 1998 Cross-Section – wpstr98.do](#)

**Sample output from Stata’s `svy` commands:**

[Complex samples estimation in Stata - sample output.log](#)

### ***5.7 How do I apply weights and correctly estimate variances in SPSS?***

Since Version 12, SPSS has included a special suite of procedures for the analysis of data from complex samples. These procedures are contained within the Complex Samples Module and are described fully in the dedicated SPSS Reference Manual on Complex Samples.

The analyst first uses the `CSPLAN ANALYSIS` command to set up an analysis plan that informs SPSS about the survey weights and the various features of the sample design. Once an analysis plan has been set up, it can then be called upon by the various estimation procedures (e.g. `CSDESCRIPTIVES`) to ensure that SPSS takes account of the features of the sample design during estimation to produce weighted estimates and correct standard errors.

The various features of the WERS sample design are specified on the `CSPLAN ANALYSIS` command in the following way:

#### ***When analysing the WERS 2004 MQ:***

```
CSPLAN ANALYSIS
  /PLAN FILE='your_path\xs04_mq.csaplan'
  /PLANVARS ANALYSISWEIGHT=estwtnr
  /PRINT PLAN
```

## 5. Weighting and statistical inference in WERS 2004

```
/DESIGN STRATA= wpstr04  
/ESTIMATOR TYPE=WR.
```

### **where:**

estwtnr is the workplace weight variable

wpstr04 identifies the stratum from which the workplace was sampled

### **When analysing the WERS 2004 ERQ:**

```
CSPLAN ANALYSIS  
/PLAN FILE='your_path\xs04_erq.csaplan'  
/PLANVARS ANALYSISWEIGHT=wrqwtmr  
/PRINT PLAN  
/DESIGN STRATA= wpstr04  
/ESTIMATOR TYPE=WR.
```

### **where:**

wrqwtmr is the employee representative weight variable

wpstr04 identifies the stratum from which the workplace was sampled

### **When analysing the WERS 2004 FPQ:**

```
CSPLAN ANALYSIS  
/PLAN FILE='your_path\xs04_fpq.csaplan'  
/PLANVARS ANALYSISWEIGHT=fpqwtmr  
/PRINT PLAN  
/DESIGN STRATA= wpstr04  
/ESTIMATOR TYPE=WR.
```

### **where:**

fpqwtmr is the workplace weight variable for the subset of cases returning  
Financial Performance Questionnaires

### **When analysing the WERS 2004 SEQ:**

```
CSPLAN ANALYSIS  
/PLAN FILE='your_path\xs04_seq.csaplan'  
/PLANVARS ANALYSISWEIGHT=seqwtmr
```

```
/PRINT PLAN  
/DESIGN STRATA= wpstr04 CLUSTER= SERNO  
/ESTIMATOR TYPE=WR.
```

**where:**

seqwtnr is the EMPLOYEE weight variable

serno is the unique workplace identifier (designating the clustering of the employee survey sample)

**When analysing the WERS 1998-2004 Panel Survey:**

```
CSPLAN ANALYSIS  
/PLAN FILE='your_path\ps9804.csaplan'  
/PLANVARS ANALYSISWEIGHT=pqwtnr  
/PRINT PLAN  
/DESIGN STRATA= wpstr98  
/ESTIMATOR TYPE=WR.
```

**where:**

pqwtnr is the weight variable for continuing workplaces

wpstr98 identifies the stratum from which the workplace was sampled in 1998

**Where do I find these commands in SPSS 's menu system?**

The Analysis Plan is set up via:

Analyze | Complex Samples | Prepare for Analysis ...

Estimation commands are accessed via:

Analyze | Complex Samples

If your Analyze menu does not contain an entry for Complex Samples, this indicates that the module is not present in your SPSS installation. Complex Samples is an add-on module that must be purchased separately from the Base system.

**Where do I get wpstr04 and wpstr98?**

## 5. Weighting and statistical inference in WERS 2004

Syntax to derive these additional variables is provided in the table at the end of this FAQ. These variables will be added to subsequent versions of the data files distributed by the UKDA.

### ***Why don't the CSPLAN commands specify without-replacement (WOR) sampling?***

It is possible to add this term, but it has little practical effect on standard errors. In our trials with Stata's `svy` suite of commands, standard errors deviated at only the 3rd decimal place when specifying a finite population correction for the MQ dataset (computed as the number of achieved observations divided by the population total, with the latter indicated in Table 2.1 in the [WERS 2004 Technical Report](#)). Use of the term is also questionable when using non-random subsets of the full data set. It might be noted that NatCen (the fieldwork company for WERS) did not use this feature when computing design effects for the WERS 2004 Technical Report.

### ***How does SPSS estimate variances in the Complex Samples Module:***

This is not specified in the manual. However, the speed of estimation suggests that variances are estimated via linearization rather than through replication methods such as the jackknife.

### ***Sample output:***

The sample output provided in the table below includes a worked example of the application of the Complex Samples procedures to run a logistic regression using the WERS 2004 MQ data. The output also indicates the impact of specifying only certain parts of the sample design. And it indicates how the Complex Samples output differs from that obtained using `WEIGHT BY` or unweighted data.

**SPSS syntax file to create wpstr04:**

[WERS 2004 Cross-Section - wpstr04.sps](#)

**SPSS syntax file to create wpstr98:**

[WERS 1998 Cross-Section – wpstr98.sps](#)

**Sample output from SPSS Complex Samples module (draft viewer file):**

[Complex samples estimation in SPSS - sample output.rtf](#)

### ***5.8 Does it matter whether one uses Stata's `svy` commands or SPSS Complex Samples module?***

Results from the two software packages do not appear to be substantively different (compare the sample output in FAQs [5.6](#) and [5.7](#)). However, Stata's `svy` suite has a much wider range of statistical procedures than SPSS's Complex Samples Module. Follow the links below for more details.

Stata's `svy` suite: <http://www.stata.com/help.cgi?survey>

SPSS Complex Samples Module: [http://www.spss.com/complex\\_samples/](http://www.spss.com/complex_samples/)

### ***5.9 What adjustments should I make if I am analysing only a subset of the full dataset or my variables have missing values?***

No additional adjustments are required for the survey weights. Each observation in the data file has its own weight value, stored in the weight variable (e.g. ESTWTNR). Correct weighting of one observation is therefore unaffected by the inclusion or exclusion of other observations. Nevertheless, this does not obviate the need to consider the impact of missing values on the representativeness of the sample. Analysts must still take the normal precautions that they would take with any sample to ensure that missing values do not bias the results by removing a non-random section of the population from the analysis. Many textbooks include discussions on the treatment of missing values (e.g. W.H. Greene (1993) *Econometric Analysis*, 2nd ed, Oxford: Macmillan, p.273-277).

## 5. Weighting and statistical inference in WERS 2004

A separate issue can arise during variance estimation in Stata when focusing on a subset of cases (e.g. private sector workplaces), if the user encounters a situation in which any stratum is represented by just one workplace [See [FAQ 5.7](#) for further details on the stratification variable]. In such instances Stata's svy suite will not run, complaining that at least one stratum is represented by a single PSU (primary sampling unit). No such problems are encountered in SPSS. The situation is remedied in Stata by using `svydes` to identify the single-PSU stratum, and then recoding WPSTR04 so that this stratum is combined with its closest neighbouring stratum. Combinations should be achieved among adjacent size categories within the same industry, where possible, on the presumption that workplaces in these strata will be more homogeneous than those grouped from different industry sectors within the same size group. The table contained in the following pdf file provides an aid to the coding of WPSTR04: [strata identification table](#).

Note that some of the strata in the variable `wpstr04` have already been collapsed to avoid problems that might otherwise be encountered during estimation on the full Cross-Section datasets. The following table provides syntax files that go one step further and group strata together to ensure that no strata have a solitary PSU for four key subsets of the data. The four subsets are private/public sector and traded/non-traded sector. Separate syntax files have been created for each of the Cross-Section surveys:

	MQ	ERQ	FPQ	SEQ
Private sector	MQ private.do	ERQ private.do	FPQ private.do	SEQ private.do
Public sector	MQ public.do	ERQ public.do	FPQ public.do	SEQ public.do
Trading sector	MQ trading.do	ERQ trading.do	FPQ trading.do	SEQ trading.do
Non-trading sector	MQ non-trading.do	ERQ non-trading.do	FPQ non-trading.do	SEQ non-trading.do

Download the [full set of syntax files](#) (zip file).

**5.10 Colleagues tell me that a fully-specified regression model doesn't need weighting – are they right?**

Strictly speaking, yes they are. But the specification of such an unweighted model is difficult to achieve in practice since the model must incorporate covariates that fully account for the sample design.

The covariates must first account for the variations in selection and response probabilities. One can attempt to check this by comparing weighted and unweighted estimates produced by the model. If there is no variation in the model coefficients, the sampling biases have been accounted for through the specification of the model.

In the employee survey, however, one must also account for the clustering of employee observations within workplaces. This may be done through the use of multilevel modelling procedures (also known as random-coefficients models) that take account of the hierarchical nature of the data.

If these two criteria are met, standard SRSWR-based inference methods can be used, enabling the user to benefit from smaller standard errors that would arise under the methods described in FAQs [5.6](#) and [5.7](#).

See Lohr S (1999) Sampling: Design and Analysis, Chapter 11 (Regression with Complex Survey Data), for a lengthier discussion.

**5.11 Why are all of the weights in WERS scaled to sum to 100 and how does this affect my analysis?**

Scaling the weights is achieved by multiplying the weight values for all observations by a single constant (i.e. scaled weight = weight \* constant). The advantage of scaling the weights so that they sum to 100 over the full sample is that the weighted base for any statistic indicates the percentage of the population of establishments (or employees for the SEQ) to which the statistic applies. This can be a handy reference point as you move through your analysis. The weighted base is indicated in Stata by the figure labelled 'Population size' which is a standard element in the `-svy-` output (see 'Subpop. size' if estimating for a sub-sample using the `subpop()` sub-command). In SPSS, the weighted base can be included in the output by specifying

## 5. *Weighting and statistical inference in WERS 2004*

the `POPSIZE` option on the `STATISTICS` sub-menu of any Complex Samples analysis command.

Scaling the weights has no effect on analyses conducted using Stata's `-svy-` commands or SPSS Complex Samples. The important factor for weighted estimators of means, percentages or regression coefficients is the ratio between the weight values of each workplace/employee. Consequently, you can alter the scaling of the weights (by choosing different values for the constant) and you will continue to get exactly the same weighted means, percentages and regression coefficients. The important factor for variance estimation is not the sum of the weights, but the number of actual observations (the unweighted base) and the nature of the sample design. Stata's `-svy-` commands and SPSS Complex Samples use this information (and ignore the sum of the weights) when estimating variances. See also [FAQ 5.12](#).

Scaling of weights is also discussed here:

<http://www.dcs.napier.ac.uk/peas/theoryweighting.htm#scaling>

### **5.12 *Why should I not use `iweights` (in Stata) or `WEIGHT BY` (in SPSS)?***

Computing statistics using `iweights` (in Stata) or `WEIGHT BY` (in SPSS) will give the correct point estimate (i.e. the correct value for the mean, proportion or regression coefficient) since the software will use a weighted estimator and thereby remove the known biases that are present in an unweighted statistic. However, the estimator will not correctly compute the variance (and thus the standard error) of the weighted statistic. This is because the software will use the standard SRSWR formula for variance estimation, amending it only to replace  $n$  (the number of observations) with  $\sum w$  (the sum of the weight values for these observations). The true variance is estimated in a much more complex way.

A comparison between the use of `iweights` and the `svy` procedures in Stata is provided in the sample output under [FAQ 5.6](#).

A comparison between the use of `WEIGHT BY` and use of the Complex Samples module in SPSS provided in the sample output under [FAQ 5.7](#)

Another simple way to show that `iweights` or `WEIGHT BY` do not correctly estimate variances is to re-scale the weights. This has no effect on the computed point estimate, but the standard error of the estimate can be manipulated at will simply by re-scaling the weights to make  $\sum w$  either very large (thus reducing the standard error) or very small (thus raising it). See also [FAQ 5.11](#).

### ***5.13 Where can I find more information about weighting and statistical inference?***

The PEAS web-site (Practical Exemplars on the Analysis of Surveys) provides a good introduction to the issues arising from complex surveys such as WERS. <http://www.napier.ac.uk/depts/fhls/peas/>

Sharon Lohr's book on sampling provides one of the most readable in-depth discussions for those wanting a more formal exposition of the statistical issues.

Lohr S (1999) Sampling: Design and Analysis, Pacific Grove, CA: Duxbury Press. ISBN 0-534-35361-4.

The standard authority on the analysis of complex surveys is a book edited by Skinner, Smith and Holt, but it is really only suitable for those who are undaunted by matrix algebra.

Skinner C, Holt D and Smith T (eds.) (1989) Analysis of Complex Surveys, Chichester: John Wiley and Sons.

### ***5.14 Where can I find more information about the revised weight for the Cross-Section Survey of Employees?***

The original weight (`SEQWTNR`) provided to users on the WERS 2004 Cross-Section Survey of Employees data file was not exactly as described in Section 7.4 of the [WERS 2004 Technical Report](#). Specifically, the adjustment for differential response rates by gender, which is discussed on p.102 of the Technical Report, was not included. Accordingly, a revised weight (`SEQWTNR2`) has been added to the dataset. A separate note outlines this issue in more detail and describes the impact on

## *5. Weighting and statistical inference in WERS 2004*

survey estimates of revising the weight to include the gender adjustment; this can be downloaded by clicking on the link below.

[Cross-Section Survey of Employees: Revisions to survey weighting](#)

## 6. Publishing your research

### *6.1 How do I acknowledge the use of the WERS 2004 data in publications?*

Users are reminded that the undertaking which is given to the Data Archive prior to receiving data from WERS 2004 requires them to acknowledge the roles of both the original depositors and the Archive in any publication, whether printed, electronic or broadcast, based wholly or in part on WERS 2004 data. The suggested wording is as follows:

"The author acknowledges the Department of Trade and Industry, the Economic and Social Research Council, the Advisory, Conciliation and Arbitration Service and the Policy Studies Institute as the originators of the 2004 Workplace Employment Relations Survey data, and the Data Archive at the University of Essex as the distributor of the data. The National Centre for Social Research was commissioned to conduct the survey fieldwork on behalf of the sponsors. None of these organisations bears any responsibility for the author's analysis and interpretations of the data."

Those using the 1998-04 Panel Survey data should replace the words '2004 Workplace Employment Relations Survey data' with '1998 Workplace Employee Relations Survey data and the 2004 Workplace Employment Relations Survey data'.

All works that use the data should also acknowledge their source by means of bibliographic citation. To ensure that such source attributions are captured for bibliographic indexes, citations should appear in either: a footnote; an endnote; or, if using the Harvard style of referencing, the reference list of publications. Those using the Harvard system of referencing should insert (Department of Trade and Industry, 2005) in the main body of the work at the point of first reference to the data. The appropriate wording to be used for the full citation is as follows:

Department of Trade and Industry (2005) Workplace Employment Relations Survey: Cross-Section, 2004 [computer file]. 1st ed. Colchester: The Data Archive [distributor], 21 December 2005. SN: 5294.

## *6. Publishing your research*

Or, if using the 1998-04 Panel Survey data file:

Department of Trade and Industry (2005) Workplace Employment Relations Survey, 2004: Panel Survey 1998-2004 [computer file], Colchester: The Data Archive [distributor], 21 December 2005. SN: 5294.

### ***6.2 How can I make others aware of my research outputs?***

We maintain a bibliography of research based on the WERS surveys, which currently includes references to over 300 research publications, including books, journal articles, working papers, mimeos, conference presentations and Masters'/Phd theses. This bibliography is one of the most effective ways of publicising your research among researchers and policy-makers who are interested in WERS-based research. It is accessed on a daily basis via our web-site.

Research outputs can be entered into an on-line database by visiting the following webpage: <http://www.wers2004.info/research/addresearch.php> and following the on-screen instructions. Please contact us if you should experience any difficulties in adding your research.

New research outputs using WERS can also be advertised by sending an email to the [WERS mailing list](#). Please include details of how one can obtain a copy of the research output, i.e. web link or publisher's contact details.