

# The 2011 Workplace Employment Relations Study: Frequently Asked Questions

Version 1.2  
7th January 2014

## Contents

<b>PART 1: OBTAINING DOCUMENTATION AND DATA.....</b>	<b>3</b>
1.1 <i>How do I obtain documentation about the 2011 WERS, such as copies of the survey questionnaires?.....</i>	3
1.2 <i>How do I obtain the 2011 WERS data to conduct my own analyses?.....</i>	3
1.3 <i>Are there any data items from the 2011 WERS that have not been deposited on the publicly available data?.....</i>	4
1.4 <i>What are the benefits of using the data held under Secure Access at the UK Data Service?.....</i>	5
1.5 <i>How can I match the WERS data to other surveys or databases?.....</i>	6
1.6 <i>Can I interrogate the 2011 WERS data without using a sophisticated statistical software package?.....</i>	7
<b>PART 2: WHAT'S NEW IN THE 2011 WERS?.....</b>	<b>8</b>
2.1 <i>Which are the new questions in the 2011 WERS and what are their origins?.....</i>	8
2.2 <i>Why did the 2011 WERS amalgamate the panel and cross-section samples?.....</i>	10
2.3 <i>What are the implications of the amalgamation of the panel and cross-section samples for my analysis?.....</i>	11
2.4 <i>What are the implications of having SEQ and WRQ data from panel workplaces in 2011?....</i>	12
2.5 <i>What is the nature of the response biases in the 2011 WERS and how have they been dealt with?.....</i>	14
<b>PART 3: FINDING YOUR WAY AROUND THE DATA FILES.....</b>	<b>18</b>
3.1 <i>What are the variable naming conventions in the 2011 WERS?.....</i>	18
3.2 <i>Do the records in the 2011 WERS data files have unique identifiers?.....</i>	19
3.3 <i>How do the 2011 WERS data files record responses from questions that permitted multiple responses?.....</i>	19
3.4 <i>How are missing values coded in the 2011 WERS?.....</i>	20
3.5 <i>What are the differences between the 2004 and 2011 survey questionnaires?.....</i>	21
<b>PART 4: MANIPULATING THE DATA FILES.....</b>	<b>22</b>
4.1 <i>How do I combine the separate 2011 WERS data files together for linked analysis?.....</i>	22
4.2 <i>How do I construct a panel dataset including data from 2004 and 2011?.....</i>	23
4.3 <i>Is it possible to construct a panel dataset spanning 1998, 2004 and 2011?.....</i>	23
<b>PART 5: DERIVED VARIABLES.....</b>	<b>25</b>
5.1 <i>How do I create a dummy variable from a multiple response set?.....</i>	25
<b>PART 6: WEIGHTING AND STATISTICAL INFERENCE IN THE 2011 WERS.....</b>	<b>26</b>
6.1 <i>What are the names of the weighting variables in the WERS data files?.....</i>	26
6.2 <i>Why should I weight my analyses?.....</i>	28
6.3 <i>How does weighting work?.....</i>	29
6.4 <i>Why will standard procedures give me incorrect standard errors?.....</i>	29
6.5 <i>What software packages allow me to account properly for the WERS sample design (both through weighting and correct estimation of standard errors)?.....</i>	31

6.6 How do I apply weights and correctly estimate variances in Stata?.....	31
6.7 How do I apply weights and correctly estimate variances in SPSS?.....	34
6.8 Does it matter whether one uses Stata’s svy commands or SPSS Complex Samples module?.....	34
6.9 What adjustments should I make if I am analysing only a subset of the full dataset or my variables have missing values? .....	34
6.10 Colleagues tell me that a fully-specified regression model doesn’t need weighting – are they right? .....	34
6.11 Why are all of the weights in WERS scaled to sum to 100 and how does this affect my analysis?.....	35
6.12 Why should I not use iweights (in Stata) or WEIGHT BY (in SPSS)?.....	36
6.13 Where can I find more information about weighting and statistical inference?.....	36
<b>PART 7: PUBLISHING YOUR RESEARCH.....</b>	<b>38</b>
7.1 How do I acknowledge the use of the 2011 WERS data in publications?.....	38
7.2 How can I make others aware of my research outputs? .....	38

## **PART 1: OBTAINING DOCUMENTATION AND DATA**

---

### **1.1 How do I obtain documentation about the 2011 WERS, such as copies of the survey questionnaires?**

Electronic versions of the survey questionnaires are provided on the [2011 Workplace Employment Relations Study webpage](#) on the GOV.UK website.

Copies of the survey questionnaires and other documentation are also available from the website of the [UK Data Service](#). The UK Data Service website also offers a short introductory note about the deposited data and a technical paper on the design and administration of the 2011 WERS. The note and the technical paper can be downloaded [here](#).

### **1.2 How do I obtain the 2011 WERS data to conduct my own analyses?**

The publicly available data files from the 2011 WERS are available from the [UK Data Service](#). The data files can be located in the [Data Catalogue](#) by entering either of the following terms in the search box:

Title: 2011 WERS

Study Number: 7226

All of the surveys in the WERS series can be found on the [Workplace Employment Relations Survey webpage](#) of the UK Data Service.

From there, users can download the deposited data files. The data files are available for download in SPSS, Stata and tab-delimited formats. Other formats may be available upon request.

Users will need to be registered in order to obtain the data files. Further information about registration is available [here](#).

All users using data for non-commercial purposes can download the data at no cost, but handling and p&p fees will apply to those requesting data on CD.

The UK Data Service website also provides open access to on-line versions of the survey questionnaires and other documentation relating to the 2011 WERS.

Interested users working in the UK can also apply to use the restricted-access version of the 2011 WERS data through [Secure Access](#) provided by the UK Data Service. The version of the data deposited under Secure Access contains additional data items and allows linkage to other datasets; further details are provided in [FAQ 1.3](#).

## *Part 1: Obtaining documentation and data*

Information about how to apply for access to the Secure Access version of the 2011 WERS data is available [here](#).

### ***1.3 Are there any data items from the 2011 WERS that have not been deposited on the publicly available data?***

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

The permanently withheld data items contain information that would directly identify a workplace or respondent, such as the name of the workplace, the organization to which it belonged and verbatim answers. An example would be **AORGNAME** in the Survey of Managers. These data items have been permanently withheld 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.

Some additional data items are only available in the restricted-access version of the data, which is accessible through [Secure Access](#) at the UK Data Service and also at the Office for National Statistics (ONS) Virtual Microdata Laboratory. These are data items which substantially raise the risk that an individual workplace could be identifiable from the data. They comprise: region identifiers; detailed industry codes below SIC Section level; and the data collected via the Financial Performance Questionnaire (FPQ).

The restricted-access version of the data also includes postcodes for all workplaces (for attaching additional geographical data or grouping workplaces into travel-to-work areas) and Inter-departmental Business Register (IDBR) reference numbers for those workplaces where managers consented to the linking of WERS data with other data sources (**MLINKDAT=Yes**). These IDBR reference numbers permit users to link data from the 2011 WERS to any other datasets which use the IDBR as a sampling frame, including the Annual Respondents Database and the Annual Survey of Hours and Earnings. The IDBR reference numbers also permit users to link data from the 2011 WERS to the Business Structure Database (the research version of the IDBR). Linking these datasets will only ever be possible through Secure Access at the UK Data Service or within the ONS Virtual Microdata Laboratory. Further details are provided in [FAQ 1.4](#).

#### **1.4 What are the benefits of using the data held under Secure Access at the UK Data Service?**

As noted in [FAQ 1.3](#), some data items from the 2011 WERS are only available on the restricted-access version of the data. Anyone wishing to utilise region identifiers, detailed industry codes below SIC Section level or data from the Financial Performance Questionnaire (FPQ) may only do so through Secure Access at the UK Data Service or at the Office for National Statistics' (ONS) Virtual Microdata Laboratory.

The restricted-access version of the data also includes Inter-departmental Business Register (IDBR) reference numbers for those workplaces where managers consented to the linking of WERS data with other data sources (**MLINKDAT=Yes**). These IDBR reference numbers permit users to link data from the 2011 WERS to other datasets which use the IDBR as a sampling frame, including the Annual Respondents Database and the Annual Survey of Hours and Earnings. The IDBR reference numbers also permit users to link data from the 2011 WERS to the Business Structure Database. See also [FAQ 1.5](#).

The Annual Respondents Database (ARD) contains firms' responses to the ONS Annual Business Inquiry (ABI) and provides information on a reporting unit's financial performance. This extends the financial performance data collected directly in WERS through the Financial Performance Questionnaire, which is also only accessible through Secure Access for the 2011 WERS. Linking data from WERS to the ARD, as well as the FPQ, can allow users to fill gaps in the FPQ data, thereby increasing the number of observations available for analysis. Another advantage is that the ARD sometimes provides more timely data.

The Annual Survey of Hours and Earnings (ASHE) is an annual survey in which employers report on the working hours and earnings of around 180,000 randomly-selected employees. The survey provides detailed information about components of wages (basic wages, overtime pay, incentive payments, pensions), which is linkable over time. 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 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.

## *Part 1: Obtaining documentation and data*

General information about Secure Access, such as how to apply to become a member, is available from the [Secure Access](#) website.

### ***1.5 How can I match the WERS data to other surveys or databases?***

In the 2,477 workplaces where managers consented to the linking of WERS data with other data sources (**MLINKDAT=Yes**), it is possible for analysts to link the WERS records to those from other surveys or administrative databases. This is subject to the requirement that the resulting linked data remains anonymised and is only used for statistical and analytical purposes.

Such data linking can be undertaken within the Virtual Microdata Laboratory operated by the Office for National Statistics (ONS) or through Secure Access at the UK Data Service. The WERS dataset deposited with Secure Access and ONS contains the Inter-Departmental Business Register (IDBR) Local Unit reference number attached to each workplace when it was sampled from the IDBR (September 2010 in the case of the refreshment sample; September 2003 in the case of the panel). The dataset also contains the IDBR reference numbers for the Reporting Unit and Enterprise to which the Local Unit belonged at the point of sampling. These IDBR reference numbers allow the WERS records to be linked to records in the ONS Business Structure Database (BSD), the Annual Respondents Database (ARD), the Annual Survey of Hours and Earnings (ASHE) and any other surveys/databases which contain IDBR reference numbers. Finally, the WERS data deposited with Secure Access and ONS also contains the postcode for each surveyed establishment, enabling analysts to attach area-level data (e.g. local unemployment rates) to the WERS records.

#### *Some notes on the practicalities of data matching:*

It should be noted that the ARD does not offer Local Unit data; WERS Local Units can only be matched to ARD Reporting Units (which typically cover whole enterprises). A discussion of the practicalities of linking WERS and the ARD can be found in [WERS 2004 Technical Paper No. 1](#).

ASHE does not contain an IDBR Local Unit identifier. However the employee records in ASHE do contain the IDBR Enterprise reference number for the employing organisation and also the postcode for the employing workplace. The combination of IDBR Enterprise reference number and postcode enables a link to be established between ASHE employees and WERS workplaces. A discussion of the practicalities can be found in [WERS 2004 Technical Paper No. 3](#).

## *Part 1: Obtaining documentation and data*

Matching area-level data onto WERS via the workplace postcode will typically be facilitated by using the National Statistics Postcode Directory, which links postcodes to broader geographical units such as Travel to Work Areas.

### ***1.6 Can I interrogate the 2011 WERS data without using a sophisticated statistical software package?***

The initial primary analyses of the survey data are presented in a First Findings report, which is available to download from the [2011 Workplace Employment Relations Study webpage](#) on the GOV.UK website. A more detailed primary analysis is presented in a full-length book [Employment Relations in the Shadow of Recession](#), published by Palgrave Macmillan. Supplementary tables to accompany the First Findings and the book are available [here](#). These tables provide further breakdowns of key data items across workplace size, industry, sector and union recognition.

The data from the 2011 WERS are also available via [Nesstar](#), the on-line analysis tool provided by the UK Data Service. 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 and WERS 2004 are also available via Nesstar.

## **PART 2: WHAT'S NEW IN THE 2011 WERS?**

---

### **2.1 Which are the new questions in the 2011 WERS and what are their origins?**

Whilst every effort is made to ensure continuity over the WERS series, changes in the world of work mean that it is important to make changes to the questionnaires. The full set of changes between the 2004 WERS and the 2011 WERS are listed in the 'Questionnaire Change Registers' which have been deposited with the survey data at the UK Data Service (SN 7226); see also [FAQ 3.5](#).

The changes included the addition of some new questions into the surveys. The most prominent additions are outlined below.

**The impact of the recession (Management Questionnaire - LRECIMP, LREACT, LRECEXP; Worker Representative Questionnaire – WGRECESS; Survey of Employees – B5):** A number of questions were inserted into the 2011 WERS to identify workplaces' and employees' experiences of the recession. The first MQ question (LRECIMP) was designed to identify the extent of the workplace's exposure to the economic downturn. The data collected by LRECIMP provide indications of the impact of the recession that are above and beyond the insights based on official statistics. LRECIMP reveals the existence of a heterogeneity in workplace's experience of the recession both within as well as between industry sectors. The second MQ question (LREACT) and the SEQ question (B5) were designed to identify the actions that were taken in response to the downturn (e.g. pay freezes, re-organisation of work). The third MQ question (LRECEXP) was designed to identify the extent to which the workplace was 'scarred' by its experience of the recession. The WRQ question (WGRECESS) was designed to allow some examination of the extent to which employees or their representatives were consulted over recession-induced changes at the workplace. The first two MQ questions (LRECIMP and LREACT) were modelled to some extent on similar questions asked by Geroski and Gregg in their study of the recession of the early 1990s - see Questions 1 and 5 in the Appendix to P Geroski and P Gregg (1997) *Coping with Recession: UK Company Performance in Adversity*, Cambridge University Press.

**Employment of non-UK nationals (EPQ - Q5):** In view of the paucity of information on the types of workplaces in which migrant workers are employed, a new question was added to the EPQ asking the workplace manager to report on the number of employees at the workplace who were not UK nationals. The Immigration (European Economic Area) Regulations 2006 gives free movement to nationals from countries within the European Economic Area (EEA), and so the workplace manager was asked to distinguish between employees from EEA countries and those from countries outside the EEA.



**Sexual orientation and religion (MQ – IPRACT; SEQ – E14, E15):** In the 2000s, equality legislation was extended to outlaw discrimination on the grounds of sexual orientation, religion and age. The 2004 MQ had asked whether sexual orientation or religious belief were among the grounds specifically mentioned in any equal opportunities policy covering the workplace (question IGROUNDS), but did not go on to ask whether the workplace's equal opportunities practices covered these grounds (as it did for gender, ethnicity, age and disability). The 2011 MQ extended the 'workplace practices' questions (IPRACT) to cover sexual orientation and religious belief. Classificatory questions were also added to the SEQ to enable more research on the in-work experiences of employees with a particular sexual orientation or religious belief.

**Provisions for mediation in workplace disputes (MQ – HRESOLX, HPROMED, HDPROMED, HMEDIATE):** The increasing focus on the internal resolution of individual workplace disputes (that is, without recourse to an Employment Tribunal) has brought with it an increasing interest in third-party mediation. The use of third-party mediation is explicitly mentioned in the Foreword to the Acas Code on Discipline and Grievance, for example. In the 2004 MQ, mediation had been mentioned only in respect of collective disputes (see GBODY). The 2011 MQ added new questions on its use to resolve individual grievances or disciplinary matters in the workplace.

**Health and safety - risks and control (MQ – IWRKRISK, IWRKCONT):** New questions were added to the MQ at the request of the Health and Safety Executive to measure the scale of risk facing employees at the workplace, and also the degree of control that employees have in managing that risk. Both questions rely on the perceptions of the workplace manager. In addition to enumerating the levels of risk and control that pertain in Britain's workplaces, it is expected that the questions may be informative in helping to explore the incidence of workplace injuries and illness (INUMINJ, ILLNUM).

**Working Time (MQ – JOPTMAN, JOPTNM; SEQ – A6):** The 2004 WERS MQ had asked about the proportion of employees at the workplace who regularly worked in excess of 48 hours per week (JEXHMAN, JEXHNMAN). However, data on the number of hours worked by employees is also available from SEQ and from other large-scale surveys such as the Labour Force Survey. In contrast, there is little information on the incidence of employee opt-outs from the Working Time Regulations. Accordingly JEXHMAN and JEXHNMAN were replaced in 2011 with two new questions asking about the proportion of employees at the workplace who had signed an opt-out agreement. A new question (A6) was also added to the SEQ to gauge employees'

## *Part 2: What's new in the 2011 WERS?*

perceptions of whether working long hours was seen as necessary to career progression within the workplace.

**Work-life conflict (SEQ – B2):** The WERS MQ and SEQ have traditionally focused on measuring the availability of practices to facilitate flexible working and work-life balance. However as a means of investigating the potential effectiveness of different types of arrangement, the 2011 SEQ added a new question to identify the degree of work-life conflict experienced by individual employees. The two-part question was modelled on questions in the 2006 British Social Attitudes Survey (CD2-30B and CD2-30D), with two principal amendments: the WERS question was bi-directional (so looking at work->non-work conflict and non-work-> work conflict); and the response scales were altered from ones based on frequency (how often do you experience conflict) to Likert scales (to what extent do you agree/disagree that your commitments in and outside work conflict with one another).

**Components of the employees' wage (SEQ – E12):** The WERS MQ has always included questions to identify the workplace's use of performance-related pay schemes. However employees' receipt of performance-related payments has never been investigated in the SEQ. A new question was added to the SEQ in 2011 to identify whether the employee received any performance-related payments as part of their wage, so that the characteristics of employees who do (and do not) receive such payments could be more fully explored. The question also identified the receipt of employer pension contributions, given the prominence of the Workplace Pension Reforms.

### ***2.2 Why did the 2011 WERS amalgamate the panel and cross-section samples?***

A major feature of the design of previous surveys in the WERS series (including the 2004 survey) was the use of two entirely separate samples. The first was a fresh cross-section sample of workplaces that were surveyed to provide representative results for the population of workplaces in existence at the time of the survey. The second was a panel sample which consisted of all participating workplaces from the previous cross-section who remained in existence at the time of the new WERS (so-called 'continuing workplaces'); the 2004 panel sample comprised all responding workplaces from the 1998 cross-section, for example. The panel sample was used primarily to assess the extent to which individual workplaces changed their behaviour over time. The full set of questionnaires was administered in the fresh sample, whilst panel workplaces only received an abridged version of the management interview. The samples were kept separate and were always analysed separately.

## *Part 2: What's new in the 2011 WERS?*

This historic design inevitably involves some duplication of effort, since the population covered by the panel survey is actually a subset of the larger population covered by the cross-section survey. This is because many of the workplaces in the current population are themselves 'continuing workplaces'. Previous surveys in the WERS series had essentially treated this overlap in the in-scope populations as a nuisance, seeking to keep the samples drawn for the cross-section and panel surveys entirely separate. The sample design for the 2011 WERS sought to take advantage of the overlap, however, by embedding the panel sample within the sample drawn for the cross-sectional survey.

This made the sample design more efficient as fewer workplaces needed to participate in the survey overall in order to achieve our twin goals of 900 panel observations and 2,700 cross-section observations. Specifically, the 2,700 cross-section observations were made up of the 900 panel workplaces plus a further 1,800 workplaces drawn from a new (and independent) refreshment sample. In the field, workplaces in both the panel and refreshment samples were each eligible for all four components of the 2011 WERS (the Management Questionnaire, Employee Representative Questionnaire, Survey of Employees and Financial Performance Questionnaire).

### ***2.3 What are the implications of the amalgamation of the panel and cross-section samples for my analysis?***

The first practical implication is that users of the survey data have access to management, employee and worker representative data for both waves of the 2004-2011 panel. Under the historic design, only management data were available in wave two of the panel. The panel data do not necessarily come from the same individuals in both years however (see [FAQ 2.4](#) for a further discussion).

The second practical implication is that the cross-sectional samples for 2004 and 2011 have some 900 workplaces in common. They are therefore not wholly independent samples and this ought to be taken into account when computing standard errors from pooled analyses that use both years of data (e.g. when testing whether a practice is more common in 2011 than in 2004). In Stata this can be done by specifying the workplace serial number (*serno*) as a 'clustering' variable on the `svyset-` command. Specifically, if one were using a pooled dataset comprising Management Questionnaire data from 2004 and 2011, one would write:

```
svyset serno [pw=estwt]
```

where *estwt=estwtncr* in the 2011 data and *estwt=estwtncr* in the 2004 data.

## *Part 2: What's new in the 2011 WERS?*

A third practical implication which, as noted below, can be ignored by most users, is that the amalgamated sample design relies on the panel and refreshment samples being equally representative of the underlying population of continuing workplaces. If the two samples are not equally representative of this underlying population, then the combined sample design will have generated different estimates for the full population of workplaces in 2011 (the cross-sectional population) than if one had obtained those estimates from a completely fresh sample (the traditional approach). This could be construed as introducing a bias into the cross-sectional estimates produced under the embedded design, and as a point of discontinuity between 2004 and 2011.

This issue was investigated by comparing the incidence of a selection of employment relations practices and workplace characteristics between the weighted combined sample and the weighted refreshment sample. The results are shown in the technical paper entitled 'The Design and Administration of the 2011 Workplace Employment Relations Survey' which accompanies the datasets released by UK Data Service. Table 4 in that document shows 54 individual estimates computed in both the combined and refreshment samples; only three exhibited differences that were statistically significant at the five per cent level (the percentage of workplaces where process operatives were the largest occupational group; the percentage operating in SIC(2007) Section L and the percentage operating in SIC(2007) Section S). None of the employment relations practices included in the examination exhibited differences that were statistically significant at the five per cent level. We therefore judge that the use of a combined sample in the 2011 WERS does not introduce a substantive point of discontinuity into the WERS cross-sectional time series.

### ***2.4 What are the implications of having SEQ and WRQ data from panel workplaces in 2011?***

The panel component of the 2011 WERS tracked down each of the 2,295 workplaces that responded to the 2004 WERS Cross-Section Survey. All of those workplaces that were still operating with 5 or more employees (so-called 'continuing establishments') were approached for interview. In previous panel surveys in the WERS series, panel workplaces had only been asked to complete an Employee Profile Questionnaire (EPQ) and Management Questionnaire (MQ) at wave two. However in 2011, panel workplaces were asked to participate in all of the components of WERS: EPQ, MQ, Worker Representative Questionnaire (WRQ), Survey of Employees Questionnaire (SEQ) and Financial Performance Questionnaire (FPQ). The 2004-2011 WERS panel sample is therefore the first in the WERS series where a subset of workplaces have WRQ, SEQ and FPQ data at two points in time.

Part 2: What's new in the 2011 WERS?

The 2004-2011 WERS panel offers the following numbers of observations:

	Total number of observations from panel workplaces		Workplaces with observations in 2004 and 2011	Number of observations in these workplaces		Panel weight
	2004	2011		2004	2011	
<b>MQ</b>	989	989	989	989	989	PQWTNR
<b>WRQ (all)</b>	512	432	281	320	318	
<b>Union</b>	394	337	253	253	253	TUWTNRP
<b>Non-union</b>	118	95	36	36	36	
<b>SEQ</b>	10,270	8,821	600	7,943	7,324	SEQWTNRP
<b>FPQ</b>	503	200	126	126	126	

Note: the FPQ was only issued to private sector workplaces in 2011

Since SEQs, WRQs and FPQs are not obtained in every panel workplace, the samples of workplaces that yield SEQs, WRQs or FPQs in each of the two waves do not perfectly overlap. Accordingly, analysts will typically want to make use of the samples indicated in the final three columns of the table above (e.g. SEQ responses from panel workplaces that generated SEQs in both 2004 and 2011). Where these 'two wave' samples are reasonably large (as is the case for the MQ, SEQ and Union WRQ), panel weights have been derived which enable the analyst to compare the responses given in 2004 with those given in 2011 in the same workplace (although it should be noted that the respondents may be different). The resulting estimates are representative of all continuing workplaces (in the case of the MQ), all employees in continuing workplaces (in the case of the SEQ) and all union reps in continuing workplaces (in the case of the Union WRQ).

The availability of multiple SEQ responses from each panel workplace means that the SEQ panel data may be analysed in two ways. The first retains the employee as the unit of analysis and compares the aggregate responses from all employees in continuing workplaces in 2004 with the aggregate responses from all employees in the same set of continuing workplaces in 2011. The data should be weighted by SEQWTNRP under this approach. However the panel sample is particularly suited to the analysis of changes in behaviour, practices or attitudes *within the same workplaces*; the second approach therefore takes the workplace as the unit of analysis, computing workplace-level summary measures from the SEQ responses (e.g. average job satisfaction) and then comparing the mean values obtained for each workplace in 2004 with the equivalent mean value obtained for the same workplace in 2011.<sup>1</sup> The workplace means should be computed on unweighted SEQ

<sup>1</sup> The workplace level summary measures can be computed using `-collapse-` or `-egen-` in Stata, or the `AGGREGATE` command in SPSS.

## *Part 2: What's new in the 2011 WERS?*

data for the sample of workplaces where SEQWTNRP is not missing; the resulting workplace-level data should be weighted by PQWTNR for analysis.

Analysts may wish to examine the sensitivity of their results to an approach which give less weight to SEQ workplace means that have been computed in establishments where the SEQ response rate ( $N \text{ SEQs returned} / \min(\text{ZALLEMPS}, 25)$ ) is low. The workplace means from establishments with a high SEQ response rate are less likely to be subject to non-response bias, although these workplaces may themselves be a non-random subset. Clearly there are trade-offs to be made between the representativeness of the sample of workplaces in which the means are computed, and the representativeness of the individual workplace means.

Note: The construction of a panel dataset from the deposited data for the 2004 WERS and the 2011 WERS is covered in [FAQ 3.2](#).

### ***2.5 What is the nature of the response biases in the 2011 WERS and how have they been dealt with?***

In common with many other voluntary surveys, WERS has seen a decline in response rates over time. For example, the overall response rate for the Management Questionnaire in 2011 was 46% (compared with 64% in 2004) and that for the Survey of Employees was 54% (compared with 61% in 2004).

Patterns of non-response were investigated for each of the instruments (Management, Worker Representatives, Survey of Employees and Financial Performance Questionnaires) and in both of the survey samples (refreshment and panel). In each case, a set of characteristics were entered (forward stepwise) into a logistic regression model in order to identify those features of the sampled unit which had a statistically significant association with the probability of obtaining a response. The resulting models were then used to generate predicted probabilities of response, the inverse of which were used to generate non-response weights. Each of the final weights provided on the deposited WERS dataset incorporates such a non-response adjustment.

The characteristics entered into the 2011 non-response analyses are listed below. Those that are underlined were found to be significantly associated with the probability of obtaining a response and were used to compute non-response adjustment factors for the instrument and sample in question.

The non-response analysis undertaken for the 2011 refreshment sample Management Questionnaire was more extensive than that undertaken for the 2004 cross-section Management Questionnaire. In order to optimise the validity of any comparisons between the 2004 cross-section and the 2011 combined sample, a

*Part 2: What's new in the 2011 WERS?*

revised set of 2004 weights has been derived using the same approach to non-response adjustment as was adopted for the 2011 data. These revised weights are available from the UK Data Service. However investigations indicate that using the revised weights will only alter the value of a small number of 2004 Management Questionnaire estimates by between one and two percentage points.

***Management Questionnaire (MQ): Refreshment sample:*** SIC(2007) industry sector; number of employees at the workplace; government office region; legal status; wave of sample allocation; whether part of a larger organisation; number of sites in the whole organisation; number of employees in the whole organisation.

***Panel sample (all measured in 2004 except \*):*** SIC(2003) industry sector; number of employees at the workplace; government office region; private or public sector; whether part of a larger organisation; number of employees in the whole organisation; largest non-managerial occupational group; whether interview at workplace or head office; whether any productive SEQs; whether any productive FPQ; whether agreed to data linking; % time spent by workplace manager on employment relations issues; any owner-manager; whether accredited as an Investor in People; any strategic plan covering employment relations; any union present; any consultative committee; any non-union representatives; any workforce meetings or briefing groups; climate of employment relations; financial performance; wave of sample allocation\*

***Worker Representative Questionnaire (union reps):*** number of employees at the workplace; SIC(2007) industry sector; legal status; region in which workplace is located; identity of largest occupational group; whether MQ completed by HR manager, general manager or finance manager; whether branch site, HQ or single independent establishment; whether MQ interview took place on site; climate of employment relations; any collective dispute in last 12 months; whether unions recognised; whether any home workers; any redundancies in last 12 months; MQ's subjective rating of workplace's financial performance; whether MQ respondent agreed to data linking (MLINKDAT); whether panel sample generated WRQs in 2004; number of WRQ interviews sought in 2011; type of union representative sought for interview; manager's rating of trust in union reps; wave of sample allocation.

***Worker Representative Questionnaire (non-union reps):*** number of employees at the workplace; SIC(2007) industry sector; legal status; region in which workplace is located; identity of largest occupational group; whether MQ completed by HR manager, general manager or finance manager; whether branch site, HQ or single independent establishment; whether MQ interview took place on site; climate of employment relations; any collective dispute in last 12 months; whether unions

*Part 2: What's new in the 2011 WERS?*

recognised; whether any home workers; any redundancies in last 12 months; MQ's subjective rating of workplace's financial performance; whether MQ respondent agreed to data linking (MLINKDAT); whether panel sample generated WRQs in 2004; number of WRQ interviews sought in 2011; type of non-union representative sought for interview; manager's rating of trust in non-union reps; wave of sample allocation.

***Survey of Employees (workplace participation):*** number of employees at the workplace; SIC(2007) industry sector; legal status; region in which workplace is located; identity of largest occupational group; whether MQ completed by HR manager, general manager or finance manager; whether branch site, HQ or single independent establishment; whether MQ interview took place on site; whether workplace has strategic plan covering ER issues; workplace recently conducted own employee survey; workplace has Investors in People accreditation; proportion of employees from ethnic minority groups; climate of employment relations; any collective dispute in last 12 months; whether unions recognised; whether any home workers; any redundancies in last 12 months; whether MQ respondent agreed to data linking (MLINKDAT); whether panel sample returned any SEQs in 2004; wave of sample allocation.

The achieved sample for the SEQ was also post-stratified by gender to match the overall distribution of employees by gender in the full population, as indicated by the full sample of EPQs.

***Financial Performance Questionnaire:*** number of employees at the workplace; SIC(2007) industry sector; legal status; region in which workplace is located; whether MQ completed by HR manager, general manager or finance manager; whether branch site, HQ or single independent establishment; whether workplace keeps records of profits; whether workplace keeps records of costs; whether workplace keeps records of sales; whether managers regularly disseminate financial information to staff/representatives; MQ's subjective rating of workplace's financial performance; whether MQ interview took place on site; whether MQ respondent agreed to data linking (MLINKDAT); whether agreed to participate in SEQ; length of MQ interview in minutes; whether panel sample returned an FPQ in 2004; wave of sample allocation.

There remains a possibility that the non-response models have not identified all of the non-response biases in the survey. In particular, the fieldwork and research teams are aware that there were notable challenges in engaging some of the larger financial services companies with the 2011 survey, perhaps because of the difficulties they were facing after the financial crisis. It is possible that the probability of response in this particular sector may have had other important (but unobserved)



*Part 2: What's new in the 2011 WERS?*

determinants besides those that were investigated in the non-response analysis. If this was the case, it may explain some of the volatility in a number of survey estimates for this sector between 2004 and 2011. The fieldwork and research teams do not believe that there are broader issues of uncorrected non-response bias elsewhere in the survey, however, given the extensive lists of characteristics that have been employed in the non-response analysis and weighting.

## **PART 3: FINDING YOUR WAY AROUND THE DATA FILES**

---

### **3.1 What are the variable naming conventions in the 2011 WERS?**

In general, each variable name has two parts. A one or two-character prefix indicates the section of the relevant questionnaire from which the variable arises. The remaining characters 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 **wers11\_mq\_general\_use** and **wers11\_mq\_restricted\_access**:*

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 **AHOWCX01** contains the first numeric response given by a particular manager to the question about changes of ownership, and **AHOWCX14** the fourteenth 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 **wers11\_wrq\_general\_use** and **wers11\_wrq\_restricted\_access**:*

Variables arising from the Worker 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 **WAREPTYP** arises from Section A of the Employee Representative questionnaire. Variables arising from multiple-response questions are labelled in the same way as in **wers11\_mq.\***

#### *Variables in **wers11\_seq\_general\_use** and **wers11\_seq\_restricted\_access**:*

Variables arising from the Survey of Employees have a two-character prefix. The first character (q) signifies that the variable is from the Survey of Employees. The second character points to the relevant section of the Survey of Employees Questionnaire. For questions that have a number of elements to them and in which the respondent is invited to provide an answer for each element or row (e.g. **A5**), letters (a, b etc.) are appended to the variable name to identify the response to each row (e.g. **qa5a** to **qa5c**). Questions inviting more than one box to be ticked (**B5**, **E7** and **E12**) yield one dichotomous variable for each of the possible responses (e.g. **qe7mul1** to **qe7mul17**). An additional variable with the same name as the question (**qe7** in this

example) indicates the number of boxes ticked by the respondent. Note that for **B5** the equivalent variable is called **qb5mul**.

Some questions in the Survey of Employees Questionnaire were not intended to elicit multiple responses but were multi-coded by a number of respondents. For questions that attracted multiple responses from more than 0.5% of respondents, the multiple responses have been made available on the dataset available from the UK Data Service. The naming convention of those variables is outlined in the section on multiple response sets below ([FAQ 3.3](#)).

### ***3.2 Do the records in the 2011 WERS data files have unique identifiers?***

Each workplace in the 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 **wers11\_mq.\*** with that from **wers11\_wrq.\*** in order to compare managers' and employee representatives' reports of the climate of employment relations at the workplace. Alternatively, one might combine information from **wers11\_mq.\*** with that from **wers11\_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 4.1](#).

In addition, each worker representative is uniquely identified by **SERNOPNUM** in the 2011 WERS data. Each employee is uniquely identified in the 2011 WERS data set by **PERSID**. However, in the 2004 WERS data, each worker representative is uniquely identified by the combination of **SERNO** and **WAREPTYP** and employees are identified by the combination of **SERNO** and **PERSID**.

### ***3.3 How do the 2011 WERS data files record responses from questions that permitted multiple responses?***

A number of questions in the Management Questionnaire and the Worker Representative Questionnaire permitted multiple responses (e.g. **CFACTORS** in the Management Questionnaire). 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

### *Part 3: Finding your way around the data files*

additional 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 Survey of Employees, a small number of questions were designed to accommodate multiple responses (i.e. **B5**, **E7** and **E12**). In the case of **B5**, which collected information on the impact of the recession on the respondent, an additional variable (**qb5mul**) was added to the data to indicate the number of responses given by each respondent. For **E7** and **E12**, the equivalent variables indicating the number of responses given are **qe7** and **qe12** respectively. For each question a set of dichotomous variables exist for each of the possible responses. For example, for **E7**, the variables **qe7mul1** to **qe7mul17** indicate whether the employee has chosen each of the available educational qualifications.

In addition, a small number of people gave multiple responses to a number of questions in the Survey of Employees despite the questionnaire indicating that the respondent should tick only one box. In those cases, the main variable is coded -6. For questions that attracted multiple responses from more than 0.5% of respondents (**D2c**, **D2d** and **D2e**), the multiple responses have been made available on the dataset adjacent to the main variable (for example, **qd2cmul1** to **qd2cmul5**). Also included is a variable which indicates the number of boxes ticked by the respondent (eg **qd2cmul**).

#### **3.4 How are missing values coded in the 2011 WERS?**

The standard approach to the coding of missing values on the 2011 WERS is to use the same codes in each data file:

- 9 - Not answered / Refused
- 8 - Don't know
- 1 - Item not applicable, i.e. not asked
- 6 - single-coded questions in the Survey of Employees Questionnaire that were inadvertently multi-coded by respondents (e.g. **qd2c**)

### *Part 3: Finding your way around the data files*

These codes are not designated as 'user missing values' in the deposited data files. Users therefore have to assign them as missing to exclude them from any statistical procedures, such as the production of means.

#### **3.5 What are the differences between the 2004 and 2011 survey questionnaires?**

Users of the 2004 WERS may be seeking to use the 2011 WERS for similar avenues of research. Some of this may be longitudinal in nature, where the research focuses on comparability of data items between the two surveys. This issue of comparability is particularly relevant for those users wishing to make use of the panel element of the 2011 WERS data. To aid these research activities, we have documented the substantive changes to the Management Questionnaire, the Worker Representative Questionnaire and the Survey of Employees Questionnaire, providing a reference guide to users who are already familiar with WERS 2004, or seeking to conduct comparative analysis.

The [Questionnaire Change Registers](#) are available from the website of the UK Data Service.

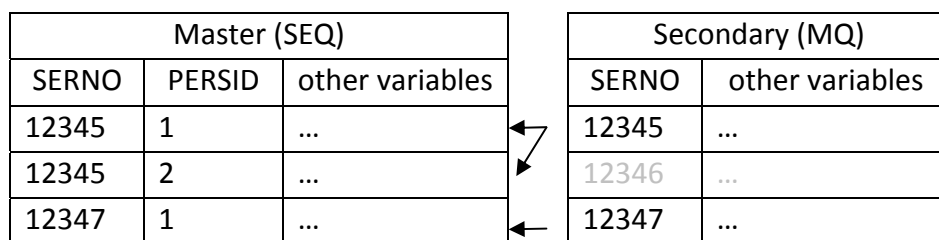
## PART 4: MANIPULATING THE DATA FILES

---

### 4.1 How do I combine the separate 2011 WERS data files together for linked analysis?

There are a number of different reasons why one might want to link the separate 2011 WERS data files together. For instance, users may wish to add data from the Management data file to the Employee Representative data file, in order to compare the responses from managers and employee representatives in the same workplaces. Alternatively, one might wish to add data from the Management data file to the Survey of Employees data file, in order to allow workplace data to be used in the analysis of employee responses. These are both ‘many-to-one’ matches, and either of these tasks can be achieved using the MATCH FILES command in SPSS or the `-merge m:1 -` 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 Management Questionnaire - MQ) being matched onto data from a master file (in this case the Survey of Employees Questionnaire - 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 specify the option `-keep(master match)-` on the `-merge-` command.

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. Summary measures at workplace level may be

#### *Part 4: Manipulating the data files*

produced from the employee data using the AGGREGATE command in SPSS or the -collapse- command in Stata. A one-to-one match can then be effected using MATCH FILES (in SPSS) or –merge 1:1- (in Stata).

#### ***4.2 How do I construct a panel dataset including data from 2004 and 2011?***

The data that comprise the 2004-2011 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 2004 Cross-Section Survey and consists of observations on 2,295 workplaces with 5 or more employees; this data file has study number 5294 in the UK Data Service's on-line catalogue. These data must be manually linked to the 2011 WERS dataset (Study Number 7226) to construct a panel dataset. Some 989 workplaces are observed in both SN 5294 and SN 7226.

The simplest way to combine the data from 2004 and 2011 is to create a 'long' dataset for each instrument (e.g. the Management Questionnaire). In this format, each workplace has two records in the data file. The first record contains data from 2004 and the second contains data from the same workplace in 2011. An index variable indicates the year to which each row relates. Derived variables may then be created which compare each workplace's answers across the two years (rows) to identify whether workplace practices have either changed or remained the same.

In creating such a dataset, users must be wary of changes in question wording or response codes between 2004 and 2011 which would render the data non-comparable. [FAQ 3.5](#) describes 'change registers' which are available from the UK Data Service and which identify such alterations in the questionnaires between the two years. Any non-comparable questions must be given distinct names in the 2004 and 2011 datasets prior to combining the two files.

In the 'long' form of the panel dataset, the panel weights (PQESTWTNR and PQEMPWTNR) which are deposited on SN 7226 should be copied onto the 2004 record, so that both observations for the same workplace have the same weighting value. The weight is invariant across time for a single workplace, and weights up the panel sample to be representative of all workplaces that had 5+ employees in 2004, 5+ employees in 2011 and which continued in operation between these two time points.

#### ***4.3 Is it possible to construct a panel dataset spanning 1998, 2004 and 2011?***

No this is not possible. Each panel survey in the WERS series is a two-wave panel only. In other words, workplaces are followed up only once and it is intended that

*Part 4: Manipulating the data files*

there is no overlap between the 1998-2004 panel sample and the 2004-2011 panel sample.



## PART 5: DERIVED VARIABLES

---

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

The example given here takes the multiple response set **CFACTORS** from the 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 and cfactor1<=9 .  
+ compute cfac_age=any(5,cfactor1 to cfactor9,xcfact1 to xcfact3) .  
end if .
```

To create the dummy variable in Stata, type:

```
egen cfac_age=eqany(cfactor1-cfactor9 xcfact1-xcfact3), values(5)  
recode cfac_age 0=. if cfactor1>9
```

The resulting dummy variable has 244 cases coded 1 and 2,432 cases coded 0. Those four respondents that did not answer **CFACTORS** remain missing on the dummy.

## **PART 6: WEIGHTING AND STATISTICAL INFERENCE IN THE 2011 WERS**

---

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.

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

The weights for the 2011 WERS are as follows:

#### **Combined sample weights:**

<i>ESTWTNRC</i>	The standard establishment weight to be applied to the cross-section combined sample of Managers.
<i>EMPWTNRC</i>	To be applied to the cross-section Survey of Managers when one wants to produce analyses reflecting the proportion of employees to whom a particular workplace characteristic pertains.
<i>SEQWTNRC</i>	Employee weight variable to be applied to the Survey of Employees.
<i>FPQWTNRC</i>	Weight to be applied to the data from the Financial Performance Questionnaire.
<i>WRQWTNRC</i>	Weight to be applied to the data for worker representatives from the Survey of Worker Representatives.
<i>PAIRWTNRC</i>	Weight to be applied to the data from the Survey of Worker Representatives when responses from union reps are being compared with responses from non-union reps in the same workplaces.

**Panel sample weights:**

<i>PQWTNR</i>	Establishment weight to be applied to the Panel Survey.
<i>PQEMPWTNR</i>	Employee weight to be applied to the Panel Survey
<i>TUWTNRP</i>	Weight to be applied to the data for union representatives when comparing responses from 2004 to those provided in 2011 within panel workplaces that generated union rep interviews at both time points.
<i>SEQWTNRP</i>	Weight to be applied to the data from the Survey of Employees when comparing employees' responses from 2004 with those provided in 2011 within panel workplaces that generated SEQs at both time points.

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 the [technical paper](#) on the Design and Administration of the 2011 WERS.

Users should note that the 2011 WERS weights were updated in the revised deposit of September 2013, as some small improvements were made to the non-response adjustment after the publication of the first edition of the First Findings in January 2013. The revisions to the weights were relatively small but users should ensure that they are using this latest version of the weights so as to obtain the most accurate estimates.

*Revised weights for WERS 2004:*

In order to optimise the validity of 2004–11 comparisons, a revised set of 2004 weights was also derived using the same approach to non-response adjustment as was adopted for the 2011 data. The impact of the revised weights is generally small and comparisons carried out by the Research Team indicate that using the revised weights will only alter the value of a small number of 2004 Management Questionnaire estimates by between one and two percentage points. However users wishing to make direct comparisons between 2004 and 2011 should use the updated weights for 2004 so that differences in weighting methodology do not compromise the comparability of their estimates. The updated weights are distinguished from the original weights by the prefix “new\_”.

## 6.2 Why should I weight my analyses?

In short, the design of the 2011 WERS sample intentionally 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. In contrast, when the sample of workplaces was drawn for the 2011 WERS, 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 intentionally 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 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 2011 WERS).

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 [technical paper](#) on the Design and Administration of the 2011 WERS. The derivation of weights is described in Section 6 of the same paper.

The theory of weighting is described more fully here:

<http://www.restore.ac.uk/PEAS/theoryweighting.php>

### **6.3 How does weighting work?**

If we take the workplace sample as an example, each workplace is assigned a weighting 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.restore.ac.uk/PEAS/theoryweighting.php>

### **6.4 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').

The 2011 WERS incorporated a number of departures from the SRSWR design (see Section 2 of the [technical paper](#) on the Design and Administration of the 2011 WERS). 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 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 (the fieldwork company for the 2011 WERS) calculated an average DEFT of 1.47 among a range of estimates from the 2011 WERS Management Questionnaire (MQ), indicating that standard formulae will underestimate the size of standard errors from the WERS MQ by 47% on average (see Section 7.1 of the 2011 WERS Technical Report, available on request from [wers@bis.gsi.gov.uk](mailto:wers@bis.gsi.gov.uk)). The average DEFT for the Survey of Employees Questionnaire was 2.03 (see Section 7.2), 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 2011 WERS 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, including Stata and SPSS, provide such methods of variance estimation.

For a fuller discussion of design effects and the correct estimation of standard errors in complex surveys, see: <http://www.restore.ac.uk/PEAS/errors.php>

**6.5 What software packages allow me to account properly 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 Reference Manual on Survey Data.

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.

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 6.12](#).

Stata and SPSS are not the only software packages to correctly account for complex sample designs.

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

Stata has 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 Reference Manual on Survey Data.

The analyst first uses the `svyset` command to inform Stata about the survey weights and various specific 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 practice, the complicated nature of the WERS sample design means that it is only possible to specify some of its features on the `svyset` command: these are the use of unequal sampling fractions; the clustering of employee observations within workplaces; and the repeated workplace observations in the panel. However these aspects of the sample design are the most important from the point of view of

determining an accurate estimate of the standard error. The impact of those other elements of the sample design listed in [FAQ 6.4](#) is typically very small indeed.

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

***When analysing the 2011 WERS Management Questionnaire:***

```
svyset [pweight=estwtnc], vce(linearized)
```

*where:*

estwtnc is the workplace weight variable for the combined sample

***When analysing the 2011 WERS Worker Representative Questionnaire:***

```
svyset [pweight=wrqwtnc], vce(linearized)
```

*where:*

wrqwtnc is the employee representative weight variable for the combined sample

***When analysing the 2011 WERS Survey of Employees Questionnaire:***

```
svyset serno [pweight=seqwtnc], vce(linearized)
```

*where:*

seqwtnc is the employee weight variable for the combined sample

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

***When analysing the 2004-2011 WERS Panel Survey:***

```
svyset serno [pweight=pqwtnc], vce(linearized)
```

*where:*

pqwtnc is the weight variable for panel workplaces

serno is the unique workplace identifier (indicating the presence of repeated observations for each panel workplace)

***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

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

The complications of the 2011 WERS combined sample design make it difficult to specify these terms. However these aspects of the sample design have little practical effect on standard errors. Experience with WERS 2004 indicates that specifying a stratification variable (identifying the sample stratum from which the workplace originated) or a finite population correction (computed as the number of achieved observations divided by the population total) affected 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 Technical Report.

***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 6.13](#)) 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 population standard deviations:***

There is no direct means of obtaining estimates of the standard deviation of a variable within the wider population. However an indirect procedure for obtaining this statistic is outlined on the Stata web-site: [obtaining population standard deviations](#).

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

The WERS Research Team are no longer using SPSS and so we do not provide up-to-date details here. However advice and syntax is still available on the web-site of the [WERS 2004 Information and Advice Service](#).

### **6.8 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](#) on the WERS 2004 Information and Advice Service website). 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/)

### **6.9 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. Each observation in the data file has its own weight value, stored in the weight variable (e.g. ESTWTNRC). 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).

### **6.10 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 than would arise under the methods described in [FAQ 6.6](#).

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

### **6.11 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 Survey of Employees) 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 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 6.12](#).

Scaling of weights is also discussed here:

<http://www.restore.ac.uk/PEAS/theoryweighting.php#scaling>

### **6.12 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 from WERS 2004 that is referred to under [FAQ 6.8](#). The same differences will arise from using WEIGHT BY instead of the Complex Samples module in SPSS.

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 6.11](#).

Finally it should be noted by SPSS users that the CROSSTABS command may also give incorrect point estimates under the WEIGHT BY approach. This is because the basic CROSSTABS command computes column or row percentages after rounding the weighted cell count and the relevant weighted marginal base to integers. The WERS weights typically sum to 100 in each sample, for the reasons outlined in [FAQ 6.11](#), and so the weighted cell counts can be small; this can lead the column or row percentage computed by CROSSTABS under the WEIGHT BY approach to be substantively different from the true value.

### **6.13 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.restore.ac.uk/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.

## **PART 7: PUBLISHING YOUR RESEARCH**

---

### ***7.1 How do I acknowledge the use of the 2011 WERS data in publications?***

Users are reminded that the undertaking which is given to the UK Data Archive prior to receiving data from the 2011 WERS requires them to acknowledge the roles of both the original depositors and the UK Data Archive in any publication, whether printed, electronic or broadcast, based wholly or in part on the 2011 WERS data. The suggested wording is as follows:

"The author acknowledges the Department for Business, Innovation and Skills, the Economic and Social Research Council, the UK Commission for Employment and Skills, the Advisory, Conciliation and Arbitration Service and the National Institute of Economic and Social Research as the originators of the 2011 Workplace Employment Relations Study data, and the UK 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."

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 for Business, Innovation and Skills, 2013) 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 for Business, Innovation and Skills, Workplace Employee Relations Survey, 2011 [computer file]. Colchester, Essex: UK Data Archive [distributor], February 2013. SN: 7226 , <http://dx.doi.org/10.5255/UKDA-SN-7226-1>

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

An on-line bibliography of research based on the WERS surveys was set up by the WERS 2004 Information and Advice Service and continues to accept new entries. The bibliography 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.

Research outputs can be entered into the on-line database by visiting the following webpage: <http://www.wers2004.info/research/addresearch.php> and following the on-screen instructions. Please contact Lucy Stokes (l.stokes@niesr.ac.uk) 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.