



Economic and Social Data Service

---

# Weighting the Social Surveys

## ESDS Government

Author: Alasdair Crockett  
UK Data Archive and Institute for Social  
and Economic Research

Updated by Reza Afkhami  
Anthony Rafferty  
Vanessa Higgins

Version: 1.8

Date: October 2008



This document is based upon the proceedings of the Weighting the Social Surveys event held by the Cathie Marsh Centre for Census and Survey Research (CCSR) at the Royal Statistical Society on Friday 12th March 2004.

The event was chaired by Jo Wathan (CCSR, University of Manchester) and the speakers were:

Ian Plewis, Institute of Education, University of London.

Jeremy Barton, Office for National Statistics.

Susan Purdon, National Centre for Social Research.

Peter Lynn, Institute for Social and Economic Research, University of Essex.

Nick Buck, Institute for Social and Economic Research, University of Essex.

The author would like to thank all the speakers. Their comments inform all sections of this report. To see the individual presentations see: <http://www.ccsr.ac.uk/esds/events/2004-03-12/slides.shtml>

# Contents

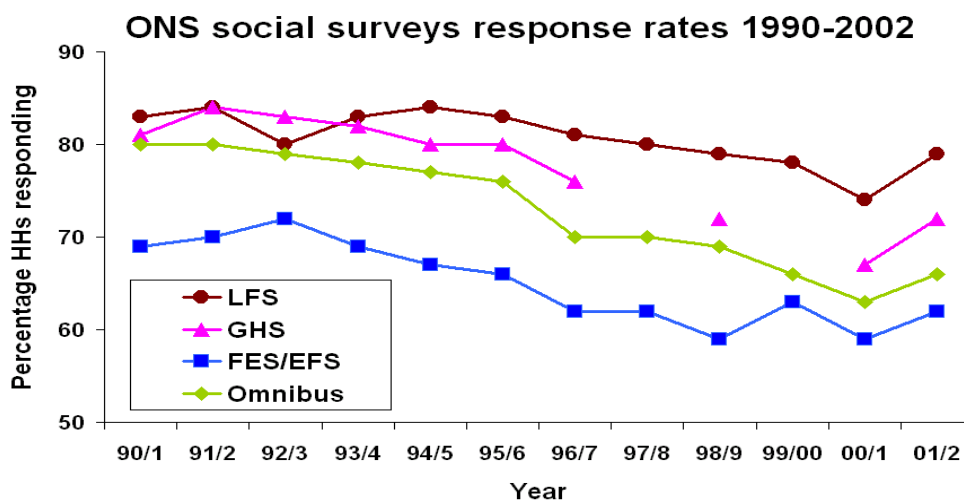
	page
1 Introduction	
1.1 What is weighting and why is it important?	5
1.2. What software to use	6
1.3 Should one always weight one's analyses?	7
1.4. What other design effects are there?	7
1.5 Illustrative case study using British Social Attitudes data	8
2 Types of weight	11
2.1 Sample design or probability weights	11
2.2 Non response weights	11
2.3 Post-stratification weights	12
2.4 Concluding remarks	13
3 Weighting Variables for the ESDS Government Surveys	14
4 References and Resources	24
5 Appendix	26

# 1. Introduction

This report aims to provide a simple guide to weighting, for users of the major government social surveys supported by the Economic and Social Data Service Government function, known as ESDS Government.

The issue of weighting, and allowing for survey design more generally, remain poorly understood by parts of the UK social science research community. The important point to realise is that unless social survey data arise from an 'equal probability of selection method' (referred to as EPSEM) and almost everyone selected agrees to be interviewed, then the sample will provide a biased representation of the total population unless adequate correction is made in subsequent analysis. This correction is usually done by weighting, to correct for the non-equal probability of selection of respondents, and differential response rates within the group of selected individuals/households. Both effects combine to mean that some types or classes of individuals are more likely to be in the achieved sample than others, and this means the achieved sample will only be representative of the population the survey aimed to reflect once the data are weighted.

In practice, few social surveys use an EPSEM design. Typically, some sample members are given a higher selection probability than others – either through a desire to over-represent important small groups in the population, or because the available sampling frame gives the researcher no choice. An example of the first type of design is when certain groups (e.g. pensioners or ethnic minorities) or regions (e.g. Northern Ireland) are deliberately over-sampled to provide more precise estimates for that group or region. An example of the second type of design is when addresses are selected using EPSEM from the Postcode Address File and then one person is selected for interview at each address (it is not possible to select persons using EPSEM as the PAF does not indicate the number of persons resident at each address). In addition, response rates are seldom close to 100% (normally less than 80%), and usually vary by the type of respondent, such that non-response is unlikely to be at random with respect to the social, demographic or economic characteristics in which the analyst is likely to be interested. The figures below shows recent trends in response rates for the Office for National Statistics major social surveys.



LFS = Labour Force Survey

GHS = General Household Survey

FES/EFS = Family Expenditure Survey/Expenditure and Food Survey

Omnibus = ONS Omnibus Survey

Source: Weighting on National Statistics Household Surveys, Jeremy Barton, Office for National Statistics, presentation given at ESDS Weighting Meeting 12<sup>th</sup> March 2004

Another common reason for weighting is to use the data for a different unit of analysis than that for which the sample was primarily designed. For example we may wish make the data representative of households as opposed to individuals, to reflect our research interests.

This report covers the issues of weighting in the government surveys supported by ESDS Government, which are generally repeated cross-sectional surveys (i.e. a different sample of people are interviewed each time the survey is conducted). Weighting is also an important, and more complex, issue in the major longitudinal surveys, in which the same group or panel of respondents are repeatedly interviewed for several years or decades (but attrition means that fewer take part each time). This report covers only surveys within the remit of the ESDS Government service. There are several important longitudinal surveys, which are covered by the ESDS Longitudinal service. Peter Lynn and Nick Buck of the Institute for Social and Economic Research were present at the meeting on which this report is based, and the slides which were produced for their excellent talks are available on the ESDS website.

## 1.1. What is weighting and why is it important?

Almost all the major British social surveys require weighting. If data requiring weighting are not weighted the resulting estimates will be biased if they are interpreted as estimates for the wider population (as opposed to estimates relating to the achieved sample). In almost all social science analysis, one is interested in the characteristics of the wider population (typically, this being the population of Britain, the United Kingdom or one or more of its constituent countries) rather than the achieved sample. For example, the British Social Attitudes Survey (BSAS) is designed to provide estimates of attitudinal data for the adult British population, but due to both differential selection probabilities (interviewing one dwelling per address, one household per dwelling<sup>1</sup>, and one adult per household), one cannot interpret the achieved sample of the BSA as providing unbiased estimates of the social attitudes of the adult British population. To generate estimates that are unbiased estimates of the British adult population, one has to weight the BSAS data.

It is important to note that the issue of bias does not just relate to complex multivariate methods such as regression, it also relates to simple descriptive statistics such as mean income, or the proportion that say they will vote at the next general election. Weighted analysis is for all social scientists, not just specialist statisticians. Indeed, for simple descriptive statistics weighting is invariably the correct thing to do, whereas for multivariate modelling there may be alternative methods that generate more precise estimates than can be achieved via weighting (see section 1.3).

A second important point to note is that weighting also involves adjustment to the precision of one's estimates. The standard measure of precision is the standard error, which tells you how close to the real value (i.e. the actual value among the population) the point or parameter estimate (e.g. means, proportions, regression coefficients) is likely to be. When data are

---

<sup>1</sup> Usually, an address is a small-user address point from the Postcode Address File. A dwelling unit is a self-contained unit of accommodation. A household is usually defined as a group of individuals who either share living accommodation or a meal per day. Definitions can vary by survey.

weighted, the precision of point and parameter estimates will tend to decline.<sup>2</sup> The precision of weighted parameter estimates will typically be lower than the corresponding precision of the unweighted estimates, though this is not always the case.

The third and final important point is that the effects of weighting are specific to each and every variable in the dataset. Some characteristics might be more common among the under-represented class(es) of respondent, others might be less common. The former characteristics will appear more common when the data are weighted, while the latter will appear less common. If the characteristics vary at random with respect to the weighting variable, the weighted and unweighted parameter estimates will be the same. Similarly, when examining relationships via regression or other techniques, the relationship between two variables might be stronger or weaker among under-represented groups. If the former, the resulting parameter estimates (e.g. regression coefficients) will be larger once the data are weighted, if the latter they will be smaller.

## 1.2 What software to use

The functionality offered by statistical software is constantly increasing. However prior to version 12, SPSS was not capable of correct weighted data analysis because it did not estimate the precision of parameter estimates correctly. This means the standard errors generated by SPSS are too small, which can lead to spurious statistical significance (as illustrated in section 1.3). The Complex sample module of version 12 of SPSS does conduct weighted analysis correctly (and also allows for design effects due to clustering and stratification), but this only covers descriptive statistics. The complex samples modules in versions beyond V.12 of SPSS contain an increasing number of multivariate commands that allow for correct survey weighting. See [http://www.spss.com/complex\\_samples/](http://www.spss.com/complex_samples/) for further details.

The other major multi-purpose statistical packages, Stata, SAS and R, all conduct weighted analysis correctly. Users of ESDS government data are advised to use Stata, SAS or R for their analyses for this reason. All ESDS government data are made available for immediate download in Stata format (as well as SPSS and tab-delimited text), however, not all procedures are available in this module.

Though not covered by this report in any detail, there are other important aspects of survey design, in addition to weighting, that affect the standard errors of survey estimates. For most of the large social surveys, these should be incorporated into one's analysis in order for standard error estimates to be correct (see 1.4 below). If you wish to incorporate these other design features, Stata, R and the specialist packages SUDAAN and WESVAR offer the greatest functionality. Stata is the most easy to use of these options and offers easy to use functionality for conducting weighted analysis and including other design features via use of the 'svy' commands. The design need only be specified once, and all subsequent commands prefixed by 'svy' will calculate standard errors in an appropriate way. The additional menu support for version 8 of Stata makes setting weights and other design features even easier. R is open source (i.e. free) package and has been ported to run in Windows as well as LINUX/UNIX, and is the best choice if your institution has no license for Stata and financial restrictions prevent you from purchasing your own license.

---

<sup>2</sup> The exception is post-stratification weights considered at section 2.3, which may increase precision. In practice, the net effect of weighting in the major social surveys is to reduce precision.

### **1.3 Should one always weight one's analyses?**

As a general principle, one should always carry out weighted analysis. If you weight by the appropriate weight variable, the point and parameter estimates you generate (e.g. means, proportions, and regression coefficients) will be unbiased population estimates. Information about weighting variables should be available in the appropriate documentation, which can be obtained from the survey pages on the ESDS webpages. The documentation should always be consulted before attempting any analysis.

Weighting to adjust for unequal sampling probabilities is therefore never a 'wrong' thing to do, but it can be sub-optimal for certain multivariate analyses in that it may reduce precision more than alternative ways of accounting for the same effects. In some models it may be possible to incorporate the information encapsulated in the weight variable (and other design features) as explicit variables on the right-hand side of one's equation. In so doing you can achieve estimates that are unbiased population estimates and may be of higher precision than would result from a weighted analysis.

If you do not weight you must incorporate all the effects encapsulated in the weighting variable, and this usually requires substantial statistical expertise. If in doubt, always weight your analyses. Incorporating design features in other ways requires specialist knowledge. If you do not know how and lack a source of expertise to ask at your institution, then you should weight your analysis; at worst this will be sub-optimal.

### **1.4 What other design effects are there?**

The effects of unequal inclusion probabilities - controlled by applying weights - are usually the most important to incorporate in one's analysis as the weighting ensures unbiased population estimates (as well as reducing the precision of estimates). Other features of survey design affect only the precision of estimates; some act to reduce precision, some to increase it. So, for statistically rigorous hypothesis testing, these design features are important. The precise nature of design effects is specific to the design of each survey. Two additional effects commonly affect British and UK social surveys; these are known as clustering and stratification effects.

#### **Clustering Effects**

Many surveys have primary sampling units (PSUs), for example post code sectors if the sampling frame is the post code address file. This means that rather than selecting the same proportion of respondents from every PSU in the population - which is very expensive and time consuming (because of the travel involved) - sample designers select a sample of PSUs and then select sample elements (e.g. households) only from the sampled PSUs. The result is that respondents are clustered within certain geographical areas. To the extent that the characteristic of interest to the researcher (e.g. income) is homogeneous within a PSU but varies between PSUs, the effect of this clustering will be to reduce the precision of population estimates.

#### **Stratification effects**

By contrast, some sample designs include stratification. Strata are groupings defined by criteria that are likely to be important to subsequent analysis, such as geographical location, social, demographic and ethnic composition, and units are sampled within these. Stratification serves to ensure that the sample is distributed over the strata in the same way as the wider population. The sample therefore better reflects the population than it would have been likely to if it were selected entirely at random. For this reason, stratification effects act to increase the precision of population estimates. The effect is stronger the stronger the relationship between the characteristic of interest to the researcher and the characteristics used to define the strata.

It is common for sample designs to incorporate both clustering and stratification elements. Each has effects on the accuracy of your results. If you ignore clustering effects (where these exist), your estimates will appear too precise - i.e. the standard errors you obtain will be under-estimates, and apparent statistical significance may be spurious as a result. Stratification effects (where these exist) act in the opposite direction though are generally weaker than clustering effects, such that clustering and stratification in combination will generally cause a modest reduction in the precision of your estimates. Information about the sample design used in your survey of interest should be available in the documentation. However, it should be noted that it will only be possible for you to obtain unbiased estimates of standard errors, taking into account the clustering and stratification, if the data set includes variables indicating PSUs and strata. Not all data sets include this information.

### **1.5 Illustrative case study using the British Social Attitudes Survey data**

Let us imagine that we are interested in changes in the rates of religious affiliation between 1994 and 2001. Let us further imagine, for the sake of simplicity, that the British Social Attitudes (BSAS) Survey was only conducted in 1994 and 2001. One then has a simple question to test from the BSAS data: was there a statistically significant difference in the proportion of British adults reporting a religious affiliation in 2001 compared with 1994.

If one examines the unweighted data the results look like this:

	Percent of BSAS respondents with a religious affiliation	Standard error
1994	62.0	0.83
2001	58.5	0.86

In terms of a formal test, the probability of this difference between 1994 and 2001 arising by chance (i.e. that both percentages arise from the same binomial distribution) is 0.004 (i.e. a 1 in 250 chance), so affiliation rates were significantly different between the two years. However, this result relates to the achieved BSAS samples, not to the adult British population. To make the BSAS estimates unbiased estimates of the adult British population; we need to apply the weight variable 'wtfactor' (a sample design weight). If we do this we get the following:

Data weighted by 'wtfactor'

	Percent of adult British population with religious affiliation	Standard error	Standard error according to SPSS
1994	61.4	0.92	0.83
2001	58.5	0.94	0.86

Note how the percentages in the second column have changed very little (so the bias of the unweighted estimate was minimal), but that the weighted standard errors in the third column are substantially higher, indicating that precision has been reduced by weighting. Note also the standard errors generated by SPSS (not using the post version 12 Complex sample module) are too small, they have not altered from the unweighted analysis shown in the previous table<sup>3</sup>.

In terms of a formal test, the probability of this difference between 1994 and 2001 arising by chance is 0.028 (i.e. the odds are 1 in 36 that the difference is due to chance).

Notice how the weighted analysis gives a much higher likelihood of a chance result (1 in 36) compared with the unweighted analysis (1 in 250). This is largely because weighting has reduced the precision of the results.

### Including the other BSAS design effects

If one adds in the clustering and stratification effects in the BSAS,<sup>4</sup> the precision of the population estimates is reduced further, as the results below show.

<sup>3</sup> In BSAS the mean weight is 1. The effective sample size in SPSS will be the same as the real achieved sample size. However, many ESDS Government surveys use weights that have a much larger mean, as the weights are used to produce population estimates. In this case, the effective sample size will be the achieved sample size times the mean weight. This will appear to reduce the standard error hugely, but this is an artefact of how SPSS applies weights in standard commands and is incorrect.

<sup>4</sup> These are a clustering effect - due to the non-equal probability of selection by PSU (post code sector), and weak stratification effects arising from the criteria used to select PSUs. In terms of using Stata, one specifies what variable corresponds to the PSU (post code sector) to account for the clustering effect, and to account for the modest stratification effect, one needs to create a new variable which is based on consecutive pairs of PSUs in terms of the order they were selected. This can be done as the data creators, the National Centre for Social Research (NATCEN), leave the variable called 'spoint' in the data supplied to ESDS government (and this gives PSU selection order by region).

Full design effect results (using Stata)

	Percent of adult British population with religious affiliation	Standard error (incorporating full design effect)
1994	61.4	1.01
2001	58.5	1.00

Note the modest additional increase in the standard errors in the third column. The probability of this difference between 1994 and 2001 arising by chance (i.e. that both percentages arise from the same binomial distribution) is now equal to 0.047, in other words there is a 1 in 21 chance that difference is due to chance.

In this case study, most of the reduction in precision arose from weighting and the additional reduction from specifying the full design effect was relatively small. This is because the BSAS design involves considerable variation in selection probabilities, while attitudes tend not to cluster greatly within postcode sectors. But this will not always be the case. For example, on a survey such as the Health Survey for England, where all persons are sampled within each sampled household (so no variation in selection probabilities), there will be no reduction in precision due to weighting (of the adult sample), whereas health measures do tend to cluster within postal sectors, resulting in a reduction in precision due to clustering.

The BSAS case study illustrates how a naïve unweighted analysis of the BSAS data would lead one to reject without hesitation the null hypothesis that a different proportion of British adults reported a religious affiliation in 2001 than in 1994. When, however, the analysis was weighted using appropriate software (Stata), and when the full design effect was specified, the difference between 1994 and 2001 was at the margins of whether we would accept or reject the null hypothesis (in both instances it would be rejected at the 0.05 significance level but accepted at the 0.01 significance level).

## 2. Types of Weight

To understand more fully what weighted analysis entails, one needs to distinguish the three primary types of weight that can exist in a given social survey dataset. These are sample design weights, non-response weights, and post-stratification weights. These three types of weight are explained in the following sub-sections.

### 2.1 Sample design or probability weights

Sample design or probability weights correct for cases having unequal probabilities of selection that result from sample design. It is important to note that non-equal selection probabilities can also occur due to differentials in non-response, which is corrected by non-response weights described at 2.2. below. Minor discrepancies may also require adjustment if the sampling frame (e.g. the postcode address file) does not entirely reflect the population, and these would constitute a type of post-stratification weight outlined at 2.3 below.

To illustrate how a sample design weight is calculated, consider a survey design that interviews one dwelling per address, one household per dwelling and one adult per household. Provided information concerning dwellings per address, households per dwelling and adults per household is enumerated by the interviewer, one can subsequently calculate sample design weights that correct for the lower selection probabilities of adults in multi-adult (and household/dwelling) households. The general formula for a sample design weight is arithmetically very simple, it is 1 divided by the probability of selection due to the survey design. However, these are usually scaled, so we define the weight as proportional to this number. For example, if there are 3 adults in a given household the resulting sample design weight for the single interviewed adult will be proportional to  $1/(1/3)$ , i.e. proportional to 3. In a one adult household, the weight will be simple proportional to  $1/1$ , i.e. proportional to one. In other words the influence of the former respondent is being increased threefold relative to the influence of the latter respondent to exactly compensate for the fact the former respondent was three times less likely to be included in the sample.

The weights are often scaled to have a mean of 1, which maintains an effective sample size when the data are weighted.

### 2.2 Non-response weights

Non-response weights compensate for differential response rates. Response rate in this sense refers to unit non-response, whereby someone refuses to take part in the survey at all, as opposed to item non-response, which relates to refusing to answer specific questions, which is addressed by missing data methods rather than weighting.

Non-response weights are typically obtained by defining weighting classes, which are based on information available for both responding and non-responding households. Such information typically relates to geographical location, primary sampling unit (PSU) characteristics (which are derived from other data sources, often the Census) and often household and dwelling type (which need to be recorded by the interviewer).

Respondents in each weighting class are weighted to compensate for the proportion of non-respondents in that class. More formally, the non-response rate weight is proportional to 1 divided by the response rate for the weighting class, i.e. directly analogous to sample design weights.

Sample design weights usually control exactly for differences in selection probability due to sample design, but non-response weights are seldom entirely accurate. The utility of the non-response weights is governed by the amount of information available to define the weighting classes. By definition, information about non-respondents is limited. The assumption of non-response weights is that the characteristics of respondents and non-respondents within each weighting class are the same; only if they are will the non-response weight be entirely accurate. If you as an analyst are examining characteristics that do vary between respondents and non-respondents within weighting classes, then the weighted estimates you derive will be biased population estimates. This problem is most likely to occur when examining measures of social engagement such as voting/not voting, which are likely to be highly correlated (even within a weighting class) with whether a respondent agrees to take part in a survey or not.

Further information on this topic is available in Ian Plewis' slides on the ESDS website (op cit).

## 2.3 Post-stratification weights

Post-stratification weights (also known as population or calibration weights) are constructed after the other types of weights have been constructed and applied to the data. They are applied to make the data even more representative of the population. As for probability weights, information on the population is usually derived from the decennial Census of Population.

These weights allow for more accurate population totals of estimates, they reduce non-response bias further (over and above non-response weights), and improve precision

Whereas sample design (probability) and non-response weights result from a very simple computation ( $1/\text{selection probability}$ ), post-stratification weights are mathematically complex, requiring iterative algorithms that maximise the fit of the data to the population. This procedure is called 'raking', and requires specialist software. The Office for National Statistics currently use a SAS based macro called CALMAR to calculate post-stratification weights, but are switching to the Generalized Estimation System (GES) programme.

For example, calculation of the Labour Force Survey individual level post-stratification weights (see: Barton 2004) involves 'raking' to 3 controls (derived from the Census and population projections):

- 5-year age group by sex within region
- Local Authority
- Single years 16-24 by sex
- Population projections

The raking procedure iterates until the data best match all three controls, and computes the post-stratification weight accordingly.

## 2.4 Concluding remarks

Few datasets supplied by the ESDS will contain distinct weight variables that correspond to these three types of weight. By multiplying the relevant weights together, one can create a single weighting variable that incorporates all three effects (or however many exist in a given study). Since this makes life easy for the secondary data user, this is typically done by the major data creators: the Office for National Statistics (ONS) and the National Centre for Social Research (NATCEN) and thereby supplied in the datasets provided by ESDS Government. The weights in these surveys will generally incorporate all the relevant weighting effects. This is not to say there is only ever one weighting variable in the dataset, there may be separate weights to make the data representative of individuals versus households, or to make the data representative of different geographical populations, e.g. the United Kingdom versus Great Britain.

While design weights are not usually changed (i.e. they remain valid in perpetuity), non-response weights and post-stratification rates may be subsequently altered to reflect new and better information becoming available to the data creator. The Office for National Statistics (ONS) is currently shifting the basis upon which weighting classes for non-response weights and post-stratification weights are calculated for surveys conducted in the late 1990s from the 1991 census (and forward projection there from) to the 2001 census (and back projection there from). As a result ONS surveys, particularly the Labour Force Surveys, are periodically resupplied to ESDS Government with recomputed weight variables. When this occurs, anyone who has previously ordered the affected datasets will automatically be notified of the resulting new edition of the data and be resupplied with the revised data and documentation.

## 3. Weighting Variables for the ESDS Government Surveys

The information given below relates to the latest available data for individual surveys. You should refer to the survey documentation on the ESDS website<sup>5</sup> for the specific year(s) you are interested in, as the weighting may change slightly from year to year.

### Annual Population Survey

This information is extracted from the APS user guide at:

<http://www.esds.ac.uk/doc/5376/mrdoc/pdf/5376userguide.pdf>

<http://www.esds.ac.uk/doc/5376/mrdoc/pdf/background.pdf>

Until December 2005 the APS comprised all Annual Population Survey Boost data (APS (B)) and LFS/LLFS data. The APS(B) covers only a subset of the topics covered on the LFS and LLFS. All variables on the LLFS appear on the APS dataset including those which are not included in the APS (B).

The main purpose of the APS weights is to gross to the population. However, this is achieved through calibration to age/sex/region totals which means that the APS weights *indirectly* deal with some of the main areas of concern for non-response.

For datasets up to and including the January – December 2005 dataset the APS requires two weighting variables due to the different data sources (the APS and the LFS) which make up the final dataset. One weight is required when looking at core variables, and one weight when looking at either only non-core variables or a combination (e.g. a crosstab) of core and non-core variables. A summary of which weight to use is as follows:

#### PWAPS04a

This is used when looking at only core variables. These are those marked as X and Y in diagram 1 in the document in the first link above.

#### PWLFS04a

This weight is used when looking at either only variables which are non-core or looking combinations of core and non-core variables. These are those marked as Z in diagram 1 in the document in the first link above.

In future the APS datasets will contain data from the LFS and LLFS and the need for two weights will disappear.

The last letter of the weighting variable changes with each quarter, as it represents the next quarter. Every quarter there will be a new weight as the weight is calculated on the sample size and characteristics. So as each new dataset is available and is different to the previous one there is a new weight calculated for each quarter and this new weight is represented by the change in last letter on the weight variable. A spreadsheet of core/non-core variables is included with the documentation from ESDS for each year of the APS (an example of this is included in the spreadsheet entitled '[weighting summary](#)' on the ESDS website).

---

<sup>5</sup> <http://www.esds.ac.uk>

## Labour Force Survey

Since 1984 the LFS has been weighted (grossed) to produce population estimates and to compensate for non-response among sub-groups. Additionally, the earnings data is also grossed. The 2006 Quarterly LFS datasets have two weights (Pwt03 and Piwt03), (1) Pwt03 is the weight for individual data - this compensates for non-response and grosses to population estimates. (2) Piwt03 is the weight for income data - this weights so that the weight of a sub-group corresponds to that sub-group's size in the population and also weights to give estimates of the number of people in certain groups. This is restricted to employees' earnings: other income data are not (yet) weighted. NB: In 2003 Pwt03 and Piwt03 replaced the weights Intiwt and Intwt because of the re-weighting exercise to bring LFS data (back to Autumn 1993) in line with the population estimates from the 2001 Census.

The QLFS household datasets contain individual level data for households, but have been designed for household analyses. They have one weight to gross to population estimates. The weight is the same for all household members. The 2005 weighting variable is called Hhwt03. See section 5 of the Household and Family Data User Guide<sup>6</sup> for more information.

The QLFS longitudinal datasets (2-quarter and 5-quarter) contain one weight to compensate for non-response and to produce population estimates. The 2005 weighting variable is called LGWT. See the Longitudinal Datasets User Guide<sup>7</sup> for more information.

Documentation for the LFS Information about the manner in which the weights have been produced including post-stratification with raking to a set of controls can be found in Barton (2004). Users should consult the survey documentation for information about the sample design, which involves a five-quarter rolling panel.

Latest userguide:

<http://www.esds.ac.uk/doc/5851%5Cmrdoc%5Cpdf%5Cbackground.pdf>

## General Household Survey

Since 2000, a dual weighting scheme has been introduced to the GHS. The dataset contains one weighting variable for two purposes (1) to compensate for non-response in the sample (2) to gross up to match known population distributions in terms of region, age-group and sex. The 2002-2002 weighting variables is called Weight01. See Appendix D in the 2002 GHS report<sup>8</sup> for more information. Full information about the methodology used for weighting the GHS can be found in Barton (2001).

### Weight variable: Weight04

The data set is unweighted. Weight04 is the variable you should use to weight the data (see details in the GHS report 'Living in Britain' or visit

<http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=5756>).

This weight applies to both household and individual level data.

See Appendix D of the [2002 GHS report](#)<sup>i</sup> or the [2004 GHS report](#)<sup>ii</sup> for more information. There is also scwght04 which is a weight for the new social capital trailer.

In February 2003, the ONS published revised mid-year population estimates for 1991 to 2000. This brought them into line with the post 2001 Census-based mid-2001 population estimates (hereafter referred to as the 2001-based intermediate population estimates). In September 2003 there was a relatively small upward revision to the mid-2001 population estimate of men (mainly

<sup>6</sup> <http://www.esds.ac.uk/doc/4820%5Cmrdoc%5Cpdf%5Chousehold.pdf>

<sup>7</sup> <http://www.esds.ac.uk/doc/5952%5Cmrdoc%5Cpdf%5Clongitudinal.pdf>

<sup>8</sup> [http://www.statistics.gov.uk/downloads/theme\\_compendia/lib2002.pdf](http://www.statistics.gov.uk/downloads/theme_compendia/lib2002.pdf)

in those aged 25 to 34 of around 190,000). There have been, and will be, further revisions to some local authority population estimates.

Latest userguide:

<http://www.esds.ac.uk/findingData/snDescription.asp?sn=5804>

## British Crime Survey

The BCS has been weighted since 1982. The survey has a number of different weights which should be applied in different circumstances. For example in the 2001 survey a variety of different weights are available for different analyses requirements. See the BCS 2001 Technical report<sup>9</sup> for a full list of weights. There are three main reasons for weighting the BCS (1) to compensate for unequal selection probabilities (2) to compensate for differential response rates (3) to ensure that quarters are equally weighted for analyses that combine data from more than one quarter. All weights include a component for unequal selection probabilities, while weighting components to compensate for differential response and to equally weight quarters are included in some weights but not in others.

Data are weighted in a number of ways for analysis. Weighting serves two purposes: to correct for different sampling rates; and to take account of 'series' of similar incidents. In the 1998 BCS, the components of the weights are:

- an **inner city weight** to correct for the over-representation of inner city residents;
- a **dwelling unit weight** to correct for cases where more than one household was at an address on the PAF file;
- an **individual weight** to correct for the under-representation of individuals living in households with more than one adult (the chance of an adult being selected for interview is inversely related to the number of adults in the household);
- a **series weight** equal to the number of incidents in the series, applied to Victim Forms representing a series of incidents.

In sweeps of the BCS which also included an ethnic boost, the boost is only included when examining results by ethnic group. The boost is excluded from all other analysis.

Three weights are derived for analysis from the components listed above:-

Weight a = individual \* inner-city \* dwelling unit

*Used for individual based analysis*

Weight b = inner-city \* dwelling unit

*Used for household based analysis*

Weight i = weight a (or weight b) \* series weight

*Used for incident based analysis*

See chapter 1 in the document below for further information

<http://www.esds.ac.uk/doc/5059/mrdoc/pdf/5059trainingguide.pdf>

See also p. 92-99 for BCS weighting information 2006/7 technical report v1

<http://www.esds.ac.uk/findingData/snDescription.asp?sn=5755#doc>

## Scottish Crime Survey

The survey has a number of different weights which should be applied in different circumstances. For example the 2000 SCS has the following weights:

- w\_house: a household weight for the main sample only. To account for (1) inaccuracies in the Postcode Address File (2) property-type bias and (3) area bias. All household data in the main sample should be analysed using this weight.
- w\_indiv: an individual weight for the main sample. The weight is a combination of household and individual weighting factors. The weight accounts for (1) different

---

<sup>9</sup> <http://www.esds.ac.uk/doc/4787/mrdoc/pdf/4787techreport.pdf>

probabilities of selection and (2) response bias towards females. All individual data in the main sample should be analysed using this weight.

- w\_person: an individual weight for the ethnic minority boost sample. This is the same as w\_indiv but without the household weighting factor. All individual data in the ethnic boost sample should be analysed using this weight.
- w\_series: a victim form series weight to reflect the fact that some victim forms refer to two or more incidents. There are two different versions on this weight: one on the main sample victim form dataset and the second on the ethnic boost sample victim form dataset. Both versions of the weight are called w\_series.

### **British Social Attitudes Survey**

The BSAS has been weighted since 1983. The 2004 survey has one sample design weight (Wtfactor) used to compensate for unequal selection probabilities (because only one person per household is interviewed). The BSAS 2004 User Guide explains this in more detail.

The datasets (in common with all surveys based on samples from the Postcode Address File) must be weighted to take account of differing selection probabilities. Simplifying slightly: households are selected with equal probability, but only one person in each household is interviewed for BSA. People in small households therefore have a higher probability of selection than people in large households and the weighting corrects for this.

*Please note that the data must be weighted in all analysis.* The file is *not* preweighted. Before running any analysis, please use the following SPSS command: weight by wtfactor.

Latest userguide:

<http://www.esds.ac.uk/findingData/snDescription.asp?sn=5823#doc>

### **Scottish Social Attitudes Survey**

The SSAS is weighted to (1) account for differing selection probabilities because only one person in the household is interviewed and (2) account for the addresses in remote and rural parts of Scotland having a greater chance of selection due to the rural boost. One weight is used (WtFactor in 2004).

Latest userguide:

<http://www.esds.ac.uk/findingData/snDescription.asp?sn=5840#doc>

### **Northern Ireland Life and Times Survey.**

All analyses of the adult data should be weighted in order to allow for disproportionate household size. In 2004 the weighting variable is called WTFactor. The only exceptions are the few household variables (for example, tenure and household income), which do not need to be weighted.

Latest userguide:

<http://www.esds.ac.uk/findingData/snDescription.asp?sn=5892#doc>

### **Young Peoples Social Attitudes**

As with the British Social Attitudes Survey (BSAS), the YPSA data were weighted to take account of the relative selection probabilities of the BSAS adult respondent at the two main stages of selection: address and household. In this respect the young people's data were weighted in the same way as the adult data. The weight on the 1998 dataset is called YPWT.

Latest userguide for 2003 survey

<http://www.esds.ac.uk/doc/5250/mrdoc/pdf/5250userguide.pdf>

## **Expenditure and Food Survey**

The EFS is weighted to adjust for non-response and to gross to population estimates. The 2002-03 dataset contains two weights: weighta and weightq. Weighta is an annual weight and weightq is a quarterly weight. The quarterly weight was introduced because sample sizes vary from quarter to quarter as a result of re-issuing addresses where there had been a non-contact or refusal to a new interviewer after an interval of a few months, so that there are more interviews in the later quarters of the year than in the first quarter. Spending patterns are seasonal and quarterly grossing counteracts any bias from the uneven spread of interviews through the year (Family Spending. A report on the 2001-02 Expenditure and Food Survey, National Statistics, Revised Edition September 2003, Appendix F, page 158). See the EFS 2001-02<sup>10</sup> report for more information.

For recent documentation please see the following link.

<http://www.esds.ac.uk/findingData/snDescription.asp?sn=5986#doc>

The data are now re-weighted to compensate for the main non-response biases identified from the 1991 Census comparison, as described in Section B6 of Family Spending 2004-05. ONS is currently undertaking a similar comparative exercise, with the 2001 Census data, which will result in an update of the non response weights.

## **Family Expenditure Survey**

Since 1998/99 the FES data has used one weight which adjusts for non-response and grosses to population estimates.

The 2000-2001 weighting variable is called "weight". Appendix F of the 2000 FES Report 'Family Spending'<sup>11</sup> contains further details of the weights.

## **Heath Survey for England**

Weighting variables are year specific owing to the variable sample design and the survey topic. For example, in 2000 weights are added for different probabilities of selection in care homes - see the 2000 User Guide<sup>12</sup>. In 2002, no weights need to be applied if only using the adult sample. If using the boost sample (on its own or together with the adult sample) a sample design weight which accounts for unequal probabilities of selection needs to be applied (tablewt).

In 2003, non-response weighting was introduced to the HSE data. Although the HSE has generally presented a good match to the population, this decision was taken to keep up with the recent changes on many large-scale government sponsored surveys, and with the aim of reducing the possible biases.

Non-response weights have been calculated for both adults and children. Four sets of non-response weights have been generated in total. Firstly a household weight was calculated to adjust for non-contact and for refusals of entire households (hhld\_wt). In addition, three sets of weights have been calculated to adjust (a) non-response among individuals in responding households, interview weight (int\_wt), (b) non-response to the nurse visit stage and (nurse\_wt), (c) refusal to give a blood sample (blood\_wt). The aim of each set of weights is that each of the main datasets (households, individuals, individuals who see a nurse, and individuals who give blood) can be treated as broadly representative of the general household population.

---

<sup>10</sup> [http://www.statistics.gov.uk/downloads/theme\\_social/Family\\_Spending\\_2001-02\\_revised/Family\\_Spending\\_revised.pdf](http://www.statistics.gov.uk/downloads/theme_social/Family_Spending_2001-02_revised/Family_Spending_revised.pdf)

<sup>11</sup> [http://www.statistics.gov.uk/downloads/theme\\_social/Family\\_Spending\\_2000-01/Family\\_Spending\\_2000-01.pdf](http://www.statistics.gov.uk/downloads/theme_social/Family_Spending_2000-01/Family_Spending_2000-01.pdf)

<sup>12</sup> <http://www.esds.ac.uk/doc/4487%5Cmrdoc%5Cpdf%5Ca4487uab.pdf>

The appropriate weight variable should be used for analysis done using data from the relevant sections. There is an extra weight (child\_wt) to compensate for limiting the number of children (aged 0-15) interviewed in a household to two. The variables int\_wt and nurse\_wt for children aged 0-15 include both the child selection weights and non- response weights. See p2 of the [2003 user guide](#)<sup>iii</sup> for more information.

Latest user guide

<http://www.esds.ac.uk/doc/6112%5Cmrdoc%5Cpdf%5C6112userguide.pdf>

## Survey of English Housing

The SEH has been weighted since 1994/95 to produce population estimates and to compensate for different response rates among households. The 2001-2002 dataset has two weight variables (H4a and H4at), both of which combine weights for non-response and grossing (1) h4a: weights for non-response and grosses to households in England (in 000s) (2) h4at: weights for non-response and grosses to tenancy groups in England (in 000s). For further information see the following document. <http://www.esds.ac.uk/doc/5021/mrdoc/pdf/5021userguide2.pdf>

There are several stages for grossing. The first is to use the sampling fraction and response rate. Broadly, if the end result of sampling and non-response is that there is an interview for one in a thousand households, the grossing factor is one thousand. The initial grossing compensates for different response rates among households that were more or less difficult to find at home, measured by the number of calls needed to make contact. Households that were harder to contact receive a bigger grossing factor than those that were easier to contact (see "Sampling fraction and response rate" below).

The remaining stages adjust the factors so that there is an exact match with population estimates, separately for males and females and for broad age groups. An important feature of the SEH grossing is that this is done by adjusting the factors for whole households, not by adjusting the factors for individuals. The population figures being matched are those for the household population and exclude people who are not covered by the SEH that is those in bed-and-breakfast accommodation, hostels, residential care homes and other institutions. There is a final stage which applies only to private tenancy groups. This compensates for the small dropout between the main stage of the survey and the private renters module.

Latest userguide 2004/5

<http://www.esds.ac.uk/findingData/snDescription.asp?sn=5709#doc>

## National Travel Survey

The NTS does not currently employ a weighting scheme, although there is consideration for the introduction of both calibration and non-response weighting schemes. However, the NTS over-samples in London due to the lower response rates achieved so there exist three variables weighted to gender, age and residency in London. These are very seldom used within the NTS and are not available from ESDS. They do not constitute a 'weighting scheme'. They are weighted versions of the main 'individual', 'number of stages' and 'number of journeys' variables. They are very minor variables, and the corresponding unweighted variables are almost always used. Details of this and the NTS sampling procedures can be found in the 2001 technical report<sup>13</sup>.

Weighting the NTS is not straightforward because of the many levels used for analysis (household, individual, vehicle, trip etc). In collaboration with NatCen, a methodology for weighting the NTS has been developed and applied to data from the 2002 NTS. This provides two sets of weights. One set, referred to as the 'diary weights', is for the sample of fully co-operating households where all members completed a travel record and the data are used for analysing trips. The other set, the 'interview weights', comprises all households which completed an interview, and therefore as well as fully co-operating households it includes 'partially responding' households, where not all individuals completed a travel record. This sample is only used for analyses that do not require travel record trip data. The weighting for both sets adjusts for household selection, household non-participation, and removal of households with missing individual interviews. Calibration weighting was carried out to adjust the weights so that the age/sex and GOR distributions of the respondents matched population estimates. This information was taken from p194-5 of the [2003-2004 Technical Report Part 1](#)<sup>iv</sup>, where there is also further information on weighting or see the [Weighting the NTS: Methodology Final Report](#)<sup>v</sup> by Pickering et al.

---

<sup>13</sup> <http://www.esds.ac.uk/findingData/snDescription.asp?sn=5340#doc>

Latest userguide: see the weighting section

<http://www.esds.ac.uk/findingData/snDescription.asp?sn=5340#doc>

Because trips of less than one mile in distance are recorded only on the seventh day of the travel week, these trips must be weighted by a factor of seven when analysed. Also for consistency with earlier surveys 'series of calls' trips are excluded from analysis of stage and trip counts and time. Tabulations, therefore, using trip or stage counts, distance or time must be performed using weighted variables. Several weighted variables have been provided to achieve this.

SXSSC	Number of stages to be counted, grossed for short walks and excluding 'series of calls' trips.
SD	Stage distance travelled, grossed for short walks.
STTXSC	Travelling time grossed for short walks and excluding 'series of calls' trips.
JJXSC	Number of trips to be counted, grossed for short walks and excluding 'series of calls' trips.
JD	Trip distance travelled, grossed for short walks.
JOTXSC	Overall trip time, grossed for short walks and excluding 'series of calls' trips.
JTTXSC	Travelling time, grossed for short walks and excluding 'series of calls'.

These weighted variables have been constructed as follows:

SSXSC

If 'series of calls'	SSXSC=0
If not 'series of calls' and 'short walk stage'	SSXSC=7
If not 'series of calls' and not 'short walk stage'	SSXSC=1

STTXSC

If 'series of calls'	STTXSC=0
If not 'series of calls' and 'short walk stage'	STTXSC=7
If not 'series of calls' and not 'short walk stage'	STTXSC=1

SD

If 'short walk stage'	SD=stage distance * 7
-----------------------	-----------------------

## National Food Survey

The weighting used in the National Food Survey is for Northern Ireland. Prior to inclusion of Northern Ireland (1996) there was no weighting. The weight accounts for the deliberate oversampling of Northern Ireland and for differential response rates among different household types. This is described in detail in the NFS User Guide<sup>14</sup>. The datasets for 1996 onwards contain an Excel file called nfsweights.xls which gives the weights that users should add to the files if using the NI data.

Weights for NFS Data 1996-2000 in the following link  
<http://www.esds.ac.uk/doc/4512/mrdoc/excel/nfsweights.xls>

## Family Resources Survey

Since 1992 the FRS used one weighting variable for two purposes (1) to gross to population (2) to compensate for non-response. However, the 1994-1995 to 2001-2002 datasets were re-released due to the inclusion of a new (interim) grossing factor introduced to make adjustments to the FRS for low income households in Scotland. These datasets contain two weighting variables: Gross1 is the original variable and Gross2 is the new variable. From 2003-04 onwards there have been further revisions to the grossing scheme - see the Grossing Review information in the FRS User Guide 1<sup>15</sup>.

Latest userguide:  
<http://www.esds.ac.uk/findingData/snDescription.asp?sn=5742#doc>

## Time Use Survey

The TUS uses weighting for a variety of reasons. There are different weights on the different files (individual questionnaire file, worksheet file, household questionnaire file and diary file). For more information go to the Time Use 2000 User Guide<sup>16</sup>.

- There are 2 individual questionnaire weights: both weights compensate for non-response and are calibrated to UK population characteristics for age-group, sex and region. The difference between the two weights is that one grosses to the UK population and the other does not. (1) wtpq Ug is the ungrossed weight which weights to the achieved sample size (2) wtpq Gr is the grossed weight which weights to UK population of those aged 8yrs or more living in private households.
- There are 2 worksheet weights: as individual weights (1) wtwrk Ug is ungrossed (2) wtwrk Gr is grossed.
- There are two diary weights: as individual weights but also compensates for differential sampling of weekdays and weekends (1) wtdwh Ug is ungrossed weight (2) wtdwh Gr is grossed.
- There are six household questionnaire weights: as individual weights but two separate weights for each of following:
  - households with dairy-keepers (1) wtdh Ug is ungrossed (2) wtdg Gr is grossed
  - households with worksheet-keepers (3) wtwh Ug is ungrossed (4) wtwg Gr is grossed
  - households with diary and worksheet-keepers (5) wtdh Ug is ungrossed (6) wtdg Gr is

grossed

## Omnibus

The Omnibus survey weights for unequal probabilities of selection caused by interviewing only one adult per household, or restricting the eligibility of the module to certain types of respondent. It should be noted that this weighting corrects for unequal probabilities of selection; it does not attempt to correct for any non-response bias.

The October 2002 dataset has two separate sample design weights (WtA and WtC) to correct for unequal probability of selection caused by either (a) interviewing only one adult per household or (b) restricting the eligibility of the module to certain types of respondent. WtA should be applied if the unit of analysis is the individual because the weight makes the sample representative of British adults.

<sup>14</sup> <http://www.esds.ac.uk/doc/4512/mrdoc/pdf/a4512uab.pdf>

<sup>15</sup> <http://www.esds.ac.uk/doc/5139/mrdoc/pdf/5139userguide1.pdf>

<sup>16</sup> <http://www.esds.ac.uk/doc/4504/mrdoc/pdf/4504userguide1.pdf>

WtC should be applied if the unit of analysis is the household reference person or spouse. Occasionally extra weights are developed separately for modules which ask questions at a different level, for example the family level.

For recent documentation see the following link.

<http://www.esds.ac.uk/findingData/snDescription.asp?sn=5813#doc>

For a copy of the Omnibus Technical Report contact the Omnibus team on [Omnibus@ons.gov.uk](mailto:Omnibus@ons.gov.uk)

## 4: References and resources

### Bibliography

#### References cited:

Barton, J. (2001) Appendix D: Living in Britain see section 4.3

Barton, J. (2004) Weighting the Social Surveys, slides from presentation, available online at <http://www.ccsr.ac.uk/esds/events/2004-03-12/slides.shtml>

#### General reading on sampling and sampling weights:

Barnett, V. (2002) *Sample Survey Principles and Methods* London: Hodder Arnold

Barton, J. (2001) 'Developing a weighting and grossing system for the GHS' *Survey methodology bulletin 49*

Butcher, B. (1984) 'Grossing Up - when and how' *Survey Methodology Bulletin 14*

Elliot, D (1991) *Weighting for Non-response: A Survey Researcher's Guide* OPCS Social Survey Division

Elliot, D (1996) 'The Presentation of Weighted Data in Survey Report Tables' *Survey Methodology Bulletin 38*

Elliot, D (1999) *Report of the Task Force on Weighting and Estimation* GSS Methodology Series 16 London: Government Statistical Service

Foster, K. (1998) *Evaluating nonresponse on Household Surveys* GSS Methodology Series 8: London Government Statistical Service

Lynn, P. (2004) 'Weighting' in Kimberly Kempf-Leonard *Encyclopedia of Social Measurement*, pp 967-974. London: Academic Press.

#### P|E|A|S (Practical Exemplars and Survey Analysis)

<http://www.dcs.napier.ac.uk/%7Epeas/about.htm>

#### ESDS Guides:

**Topic Guides:** ESDS Government produces an annual topic-oriented guide to the major cross-sectional surveys. In 2003 this was based on Employment and the Labour Market. The guide contains a summary of weighting schemes used in the surveys and clickable links to relevant documentation for individual surveys. This is available on the ESDS webpages from :

<http://www.esds.ac.uk/government/docs/> .

**Other key documents:** This page also contains a link to the GSS' 1999 Report of the Taskforce on Weighting and Estimation. The appendix of this document reviews contemporary weighting schema for a range of surveys.

#### Survey specific resources

All surveys have documentation available. This should be obtained with the data and consulted before the using the datasets.

**General Household Survey:** Appendix D of the 2001 'Living in Britain' report contains guidance on how weights have been produced for the GHS, and their effect on results. This can be found at <http://www.statistics.gov.uk/lib2001/resources/fileAttachments/GHS2001.pdf> .

A description of the methodological work undertaken to produce the weights is available in Social Survey Methodology Bulletin (SMB). 'Developing a weighting and grossing system for the General Household Survey' by Jeremy Barton was published in SMB 49 pp15ff in July 2001 and is available online at:

<http://www.statistics.gov.uk/ssd/ssmb/ssmb49.pdf> .

Appendix D of the 2004 'General Household survey' report contains guidance on how weights have been produced for the GHS, and their effect on results. This can be found at

[http://www.statistics.gov.uk/downloads/theme\\_compendia/05\\_Appendix\\_D.pdf](http://www.statistics.gov.uk/downloads/theme_compendia/05_Appendix_D.pdf)

**Labour Force Survey:** Weights are available separately for different purposes including Individual analyses and Income on the QLFS general file, Household level analyses in the Household file and for users of the longitudinal data. Information on these are available in the appropriate documentation available from the UK Data Archive. The most recent of these can be found at the following locations:

- QLFS (Individual and Income data):  
<http://www.esds.ac.uk/doc/5851%5Cmrdoc%5Cpdf%5Cbackground.pdf>
- Household data:  
<http://www.esds.ac.uk/doc/5716%5Cmrdoc%5Cpdf%5Cbackground.pdf>  
<http://www.esds.ac.uk/doc/5716%5Cmrdoc%5Cpdf%5Chousehold.pdf>
- Longitudinal data:  
<http://www.esds.ac.uk/doc/5853%5Cmrdoc%5Cpdf%5Cbackground.pdf>  
<http://www.esds.ac.uk/doc/5853%5Cmrdoc%5Cpdf%5Clongitudinal.pdf>

Guidance on the effect of regrossing in the light of updated population estimates is available in the August 2006 edition of Labour Market Trends at:

[http://www.statistics.gov.uk/downloads/theme\\_labour/LMT\\_Aug06.pdf](http://www.statistics.gov.uk/downloads/theme_labour/LMT_Aug06.pdf)

**Expenditure and Food Survey:** A description of the weighting scheme used in the EFS is available in Appendix B6 of 'Family Spending 2005 edition' a report on the 2004-5 EFS. This is available online at:

[http://www.statistics.gov.uk/downloads/theme\\_social/Family\\_Spending\\_2004-05/FS04-05.pdf](http://www.statistics.gov.uk/downloads/theme_social/Family_Spending_2004-05/FS04-05.pdf)

<http://www.esds.ac.uk/doc/5375/mrdoc/pdf/5375userguide1.pdf>

## Appendix

### Weighting Data in SPSS\*

The WEIGHT command simulates case replication by treating each case as if it were actually the number of cases indicated by the value of the weight variable. You can use a weight variable to adjust the distribution of cases to more accurately reflect the larger population or to simulate raw data from aggregated data.

#### Example

A sample data file contains 52% males and 48% females, but you know that in the larger population the real distribution is 49% males and 51% females. You can compute and apply a weight variable to simulate this distribution.

```
*weight_sample.sps.
***create sample data of 52 males, 48 females***.
NEW FILE.
INPUT PROGRAM.
- STRING gender (A6).
- LOOP #I =1 TO 100.
- DO IF #I <= 52.
- COMPUTE gender='Male'.
- ELSE.
- COMPUTE Gender='Female'.
- END IF.
- COMPUTE AgeCategory = trunc(uniform(3)+1).
- END CASE.
- END LOOP.
- END FILE.
END INPUT PROGRAM.
FREQUENCIES VARIABLES=gender AgeCategory.
***create and apply weightvar***.
***to simulate 49 males, 51 females***.
DO IF gender = 'Male'.
- COMPUTE weightvar=49/52.
ELSE IF gender = 'Female'.
- COMPUTE weightvar=51/48.
END IF.
WEIGHT BY weightvar.
FREQUENCIES VARIABLES=gender AgeCategory.
```

- Everything prior to the first FREQUENCIES command simply generates a sample dataset with 52 males and 48 females.

#### File Operations

- The DO IF structure sets one value of *weightvar* for males and a different value for females. The formula used here is: *desired proportion/observed proportion*. For males, it is 49/52 (0.94), and for females, it is 51/48 (1.06).
- The WEIGHT command weights cases by the value of *weightvar*, and the second FREQUENCIES command displays the weighted distribution.

*Note:* In this example, the weight values have been calculated in a manner that does not alter the total number of cases. If the weighted number of cases exceeds the original number of cases, tests of significance are inflated; if it is smaller, they are deflated. More flexible and reliable weighting techniques are available in the Complex Samples add-on module.

### Example

You want to calculate measures of association and/or significance tests for a crosstabulation, but all you have to work with is the summary table, not the raw data used to construct the table. The table looks like this:

	Male	Female	Total
Under \$50K	25	35	60
\$50K+	30	10	40
Total	55	45	100

You then read the data into SPSS, using rows, columns, and cell counts as variables; then, use the cell count variable as a weight variable.

```
*weight.sps.
```

```
DATA LIST LIST /Income Gender count.
```

```
BEGIN DATA
```

```
1, 1, 25
```

```
1, 2, 35
```

```
2, 1, 30
```

```
2, 2, 10
```

```
END DATA.
```

```
VALUE LABELS
```

```
Income 1 'Under $50K' 2 '$50K+'
```

```
/Gender 1 'Male' 2 'Female'.
```

```
WEIGHT BY count.
```

```
CROSSTABS TABLES=Income by Gender
```

```
/STATISTICS=CC PHI.
```

- The values for *Income* and *Gender* represent the row and column positions from the original table, and *count* is the value that appears in the corresponding cell in the table. For example, 1, 2, 35 indicate that the value in the first row, second column is 35. (The *Total* row and column are not included.)
- The VALUE LABELS command assigns descriptive labels to the numeric codes for *Income* and *Gender*. In this example, the value labels are the row and column labels from the original table.
- The WEIGHT command weights cases by the value of *count*, which is the number of cases in each cell of the original table.
- The CROSSTABS command produces a table very similar to the original and provides statistical tests of association and significance.

*Crosstabulation and significance tests for reconstructed table*

**Income \* Gender Crosstabulation**

		Gender		Total
		Male	Female	
Income	Under \$50K	25	35	60
	\$50K+	30	10	40
Total		55	45	100

**Symmetric Measures**

		Value	Approx. Sig.
Nominal by	Phi	-.328	.001
Nominal	Cramer's V	.328	.001
	Contingency Coefficient	.312	.001
N of Valid Cases		100	

\*\* This is extracted from Chapter 4 p.83-84 SPSS Programming and Data Management, 3rd Edition A Guide for SPSS and SAS® Users Raynald Levesque and SPSS Inc.

<sup>i</sup> [http://www.statistics.gov.uk/downloads/theme\\_compendia/lib2002.pdf](http://www.statistics.gov.uk/downloads/theme_compendia/lib2002.pdf)

<sup>ii</sup> [http://www.statistics.gov.uk/downloads/theme\\_compendia/05\\_Appendix\\_D.pdf](http://www.statistics.gov.uk/downloads/theme_compendia/05_Appendix_D.pdf)

<sup>iii</sup> <http://www.esds.ac.uk/doc/5098%5Cmrdoc%5Cpdf%5C5098userguide.pdf>

<sup>iv</sup> <http://www.esds.ac.uk/doc/5340/mrdoc/pdf/5340userguide1.pdf>

<sup>v</sup> [http://www.dft.gov.uk/stellent/groups/dft\\_control/documents/contentservertemplate/dft\\_index.hcst?n=14562&l=4](http://www.dft.gov.uk/stellent/groups/dft_control/documents/contentservertemplate/dft_index.hcst?n=14562&l=4)



Economic and Social Data Service

ESDS Government  
Economic and Social Data Service  
Cathie Marsh Centre for Census and Survey Research  
University of Manchester  
Manchester M13 9PL

Email: [govsurveys@esds.ac.uk](mailto:govsurveys@esds.ac.uk)  
Tel: +44 (0)161 275 1980  
Fax: 0161 275 4722  
[www.esds.ac.uk/government](http://www.esds.ac.uk/government)

