Quick Guide for Users

Dataset:	New Earnings Survey (NES)
Dates available:	1986-2002, annual
Source:	Businesses, 1% sample of employees
Coverage:	Earnings and hours of employees paying NI
Collected by:	ONS
Link fields:	PAYE reference (link to IDBR reference supplied by ONS Newport)
Legal restrictions:	Covered by Statistics of Trade Act and Data Protection Act

Quick summary:

The NES is an employer survey into the hours and earnings of employees. The sample is taken in April of the relevant year and relates to employment over the previous year. A survey form is sent to employers, and completion is compulsory under the STA. Some large organisations make automatic submissions direct from their electronic records.

It is intended to include 1% of the workforce as sampling is done by taking records with a specific final two digits on the employees NI number. It therefore also constitutes a panel, and there is in fact a New Earnings Survey Panel Dataset (NESPD), which is an anonymised and reduced form of the full set of variables (not currently held by BDL).

The survey provides a large amount of information on earnings and hours (including bonuses, overtime etc) and has industry information derived from the IDBR, but has no personal characteristics of the employee apart from age and sex.

Most variables are collected each year. In addition, each year has a few additional questions, which may or may not be asked in other years.

The sample reference is the NI number, and therefore the sampling frame is supplied by (mainly) Inland Revenue records. The NI number is linked to a company PAYE reference, which in turn can be linked to an IDBR reference (although not with a 100% success rate). ONS Newport has provided a link file with IDBR-PAYE references but the match is imperfect, especially for earlier years. The link can only be made for the IDBR years (1994 onwards); hence BDL currently only has linked files for 1994-2002, although the cleaned files are available back to 1986.

Sampling frame:

The same individuals, with the same last two digits on their NI number, are sampled every year. The NES therefore constitutes a true longitudinal study.

	Number of Observations	
Year	Original files	Cleaned files
1986	166405	166303
1987	166026	165848
1988	175824	175717
1989	174790	174650
1990	177114	176787
1991	176107	175970
1992	170880	170315
1993	163600	163534
1994	166657	166622
1995	162173	162053
1996	163836	163810
1997	155913	154069
1998	161648	161612
1999	162013	161954
2000	159149	159111
2001	161573	161516
2002	162868	162806

Organisation of files:

Original data files are stored in the admin drive, because they include NINos. These are removed as the last stage of the cleaning process and replaced by a consistent index scheme. See the file *NINOs in the NES.doc* for more information on NINos.

For some years, data is taken from two files to maximise the available variables. All years then have the variables renamed so that the repeated variables have a consistent name structure. These are the variables named in the *variable_reference.xls* spreadsheet (in this folder there is a link to the original spreadsheet from which this information was derived). The files are then cleaned, mainly to remove duplicates, and then linked, so that PAYE reference numbers are matched to IDBR references.

Known data issues:

There are known inconsistencies between years on the age and sex of employees. Some of these are feasible (e.g. a two-year gap on age instead of a one-year gap could be due to a birthday in April on or near the survey date). Some are not likely (e.g. multiple sex changes). Some are not feasible (age wrong by several years) and must be attributable to an error, but without independent information the "correct" value cannot be identified. The NESPD group at ONS has spent some time trying to correct this to create a true panel.

Some individuals have multiple records within a year. This may be legitimate – ONS telephone follow-up checks reveal some people to legitimately have several jobs. However, age and sex inconsistencies appear. There are also duplicates, some of which are easily identifiable and others not. The cleaned files have had the obvious duplicates removed. For further details see the cleaning and linking audit documents.

The survey should be limited to those earning over the NI limit. In practice, a small number of individuals earning less than that are included. This may be due to temporary dips in salary, or may be the result of automatic submission of payrolls by large companies.

The NI number is held as six characters until 1998; thereafter it is held as follows:

	6 digits	**7 digits**	**8 digits**	**9 digits**
1999	2075	0	723	159156
2000	687	0	607	157817
2001	3	1	628	160884
2002	0	0	587	162219

In theory, six characters are sufficient: the last two numbers are the (known) sampling criterion to achieve a 1% sample, and the final character is a check digit. However, in the years where we have full infomration there are a very small number of cases – tens or hundreds - where there appear to be different check digits. An examination of the data suggests that this is usually spurious but there do appear to be a few cases - tens each year – where the data suggests two different people. This implies that in the earlier yeasrs with only 6 digits there is a possibility of treating two people as the same person with two jobs; and in later years, there is the danger of treating one individual with two jobs and a wrong check digit as two different people.

The 1997 dataset contains no PAYE reference and so cannot be linked to the ARD except by pattern-matching on address. The IDBR lookup table which is used to match PAYE references to the IDBR was supplied only for the year 1997. This means that the matching that has been done is only for the snapshot of the IDBR in 1997, and so there will not be a one-to-one match between the two references. As a quick test, we have checked how many of the PAYE references do match up, and not surprisingly, the highest number of matches occur in 1997 with the numbers increasingly falling as you move away either side from 1997.

Because the NES has been running for a long time, some variable definitions have changed. In particular, the collective agreement variable (AGT) should be treated with caution from the mid-1990s.

It is worth pointing out that the “same job” variable relates to whether an individual has been in the same job for the last year, not the same company. The latter is much more common in other datasets and confusion often arises.

Data on annual earnings is only available from 1995. Data for 1995 was subject to quality problems and should not be used. For safety reasons only data from 1998 onwards should be used (as 1998 was the first year that annual earnings were validated properly and published).

One should also note that some Nino’s are temporary National Insurance numbers. These are assigned to individuals on a yearly basis and the numbers can be re-used for **different** individuals from year to year. Therefore temporary NINos (indexes marked with a T) should be ignored in the linked files since the data will relate to different individuals in different years.

Other issues:

The NES sampling frame will be changed for the 2004 survey onwards to reflect a wider distribution of pay.