Technical report

Methodology for the 2004 Annual Survey of Hours and Earnings

By **Derek Bird**, Employment Earnings and Productivity Division, Office for National Statistics

Key points

- The Annual Survey of Hours and Earnings (ASHE) replaced the New Earnings Survey (NES) from 28 October 2004.
- The ASHE samples from the PAYE system, but weights responses to the number of jobs from the Labour Force Survey.
- The ASHE sample has been increased to include employees in businesses outside of the PAYE system and employees who changed or started new jobs after sample identification.
- Imputation for item non-response has been introduced.
- The survey questionnaire has been redesigned and tested ready for introduction in 2005.
- The main publication now covers the UK, includes quality measures and has an improved layout and content.
- Results using ASHE methodology applied to NES data for 1998 to 2003 are available.

Introduction

ONS is undergoing a significant modernisation programme of its statistical systems to make them world class in the 21st century. The objectives of this Statistical Modernisation Programme (SMP) are to standardise and systematise the processing and presentation of statistical outputs.

The development of a new annual earnings survey, the ASHE, to replace the New Earnings Survey (NES) is ONS's first major survey redesign as part of this modernisation programme. The NES was designed to meet the policy needs of the 1970s and has changed little over the past 30 years. The ASHE provides an opportunity to meet users' requirements, to improve the methodology of the survey and to make use of new statistical tools.

The methodology that underpinned the annual NES has been changed in line with recommendations made in the National Statistics Quality Review of the Distribution of Earnings Statistics (DOER). The changes address the weaknesses in the NES's design, which led to the production of biased estimates of earnings. The biases arose because the survey responses to the NES were not weighted to the population of employees. Additionally, the sample yielded incomplete coverage of employees, primarily because the main source for the NES sample was the Inland Revenue's PAYE system. Other biases occurred because of differential non-response for employees of different types. Finally, the survey missed significant numbers of employees that change job between sample selection and the survey reference date, but who remain within scope of the survey since they remain in employment.

As well as addressing the weaknesses in the survey methodology, the questionnaire has also been reviewed. The NES questionnaire was poorly designed and allowed too much latitude for contributors to interpret the response requirement in their own way, which increases variation in the ▶

data. This has led to the design of a new questionnaire, which was tested on a sample of 5,000 employees alongside the 2004 ASHE survey. The parallel test allows a comparison to be made between the old and new questionnaires, to compare response rates and to test the processing system. Subject to the outcome of this field test, the new questionnaire is likely to be introduced for the 2005 ASHE.

The introduction of the new survey methodology will introduce discontinuities to statistics of earnings, but historical results using a consistent approach have been constructed to allow users to assess the impact of these changes over a reasonably long time frame. Historical results will be published on the National Statistics website for the period 1992 through 2003, though initially resource constraints mean that estimates for 1998 through 2003 were released in the first half of October to allow users to understand the impact of the improvements. These estimates were compiled by applying the ASHE methodology to the NES datasets for 1992 to 2003. An analysis of the impact of these changes was published in a separate article on the website that will be reproduced in the next issue of *Labour Market Trends*.

To generate these historical estimates ONS has created an occupational code consistent with Standard Occupational Classification (SOC) 2000 for the years 2001 back to 1992. This was done by using the NES 2002 dataset that was dual-coded to both SOC90 and SOC2000. Where employees had not changed jobs in a year the SOC2000 code was taken back. For employees that had changed jobs, a SOC2000 code was estimated using their SOC 1990 code, adjusted by using information from the dual-coding from the 2002 NES. As part of this process, the LFS calibration totals were also adjusted so that they were on an equivalent SOC2000 basis, for 2000 back to 1992.

This article takes the following form: the first section deals with issues around the weighting methodology used for the survey in 2004. This methodology produces weighted estimates of earnings, the weights are calculated by calibrating the survey responses to totals from the Labour Force Survey (LFS) by occupation, gender, region and age. The second section looks at the pilot surveys that have been conducted to assess the degree to which the inclusion of different types of employees that are currently outside the NES sample frame is likely to improve the survey results. The third section considers the redesign of the survey questionnaire and then goes on to look at the new criteria underpinning results publication and an intention to focus the survey outputs on the median in preference to the mean.

Methodology overview

The main sample file underpinning the ASHE will remain the same as for the NES. This comprises all jobs in which an employee's National Insurance number (NINo) ends with a specified pair of digits. It is obtained from Inland Revenue (IR), and is a 1 in 100 random sample of all jobs registered in a PAYE scheme. Because the main sample file includes only those jobs registered in a PAYE scheme there is an issue of undercoverage of the labour market, especially of the lower earners. This is because many of those not registered in a PAYE scheme can be expected to earn below the tax threshold. To address this issue supplementary surveys are conducted to augment the data inputs to the ASHE. As with the NES, the ASHE questionnaires collect information about employees; they are sent to employers who supply the requested employee information.

The new survey delivers weighted estimates of pay, whereas the NES delivered only unweighted ones. In order to calculate weights, responses are divided into calibration groups defined by a cross-classification of occupation, sex, age and workplace region where:

- occupation is the Standard Occupational Classification (SOC) 2000 one-digit (or major group) code, of which there are 9;
- age is split into three age bands (16-21, 22-49 and 50 and over); and
- workplace region is based upon government office region (GOR), but aggregated into two areas comprising (i) London and the South East and (ii) elsewhere in the United Kingdom.

The total number of employee first and second jobs in the LFS is used to provide calibration totals for the 108 groups (or strata). Estimates of pay and associated standard errors for different subsets of the population have been made using weighted estimation.

Forming strata

Initial work on forming strata focused on determining which variables were best associated with pay. Finding these meant that strata could be defined that would form the basis of the weighting structure. The NES 2000 response file was used for this analysis, and both hourly and weekly pay were examined. Statistical techniques were used to identify the variables for inclusion in the model.

At the outset a wide range of possible prediction variables were tested for inclusion including gender, age, occupation, place of work, industrial classification, full-time/part-time markers and so on. Several of these variables can be grouped in different ways, and these were investigated too.

Many different combinations of variables were tried, but the final decision on how the strata should be formed was not determined by the statistical analysis alone, as other issues also had an influence. The outcome needed to avoid the generation of a very large number of strata, as LFS estimated totals in smaller strata would be more subject to statistical error themselves. It was also desirable to include a number of different variables in the stratification, especially those groups that are most prominent in the publication of the survey – sex, for example.

The analyses showed that occupation is by far the best single predictor of pay. Combinations of other variables with occupation were tried, and although some others explained earnings relatively well on their own, they were found to be superfluous when combined with occupation. In the end, the decision to use SOC major group, gender, three age bands and two regions provided the best trade-off between the prediction of earnings and an excessive number of small strata.

Age was grouped into bands to make stratification easier. The age bands were formed so as to keep the closest homogeneity of average pay within groups but the groups different from each other. A secondary but important consideration was to retain a degree of correspondence with national minimum wage legislation. A check of average pay levels by age in years showed clearly how the bands should be defined. The workplace regions (defined by GOR) were aggregated into two groups: London and the South East, and elsewhere, formed on the same basis as the age bands. It is worth stating here that these strata have been defined only for weighting purposes. Domain estimation will allow estimates to be derived for any subset of the population, even if these sit within, or even span parts of, different calibration groups.

Calibration and weighting

The ASHE has 108 calibration groups or poststrata, and uses LFS estimates of employee totals (including second jobs) as calibration totals for these poststrata.

Since LFS totals are themselves estimates, ONS has analysed the sensitivity of estimates of pay when different LFS totals are used for calibration. The 2002 NES response file was calibrated to each of the LFS quarterly datasets from 2001, from 2002, and the 2001/2 annual LFS dataset, together with datasets derived as a combination (weighted means, medians, etc.) of some of the quarterly datasets. Naturally, using different LFS datasets as calibration totals results in different NES estimates of pay; the investigations were intended to allow ONS to gauge the size of these differences. The following conclusions were reached.

- The total of employee first and second jobs from the LFS will be used, as this most closely matches the ASHE, which measures jobs. The number of employees on the ASHE files with three or more jobs is small.
- The estimates of pay for large subgroups of the population, for example all employees, are relatively robust to the use of different calibration totals. The range in estimates of gross weekly pay (caused by using different LFS datasets for calibration) was about 1 per cent of the estimate itself. This is small, and the ASHE methodology uses, as calibration totals, the LFS estimates directly. The standard error of the LFS estimates themselves is small, and has not been included in the calculation of the estimates of the standard error estimates of estimators of pay.
- Ideally the annual LFS dataset would be used for calibration, as its sample size is larger than that of the quarterly datasets, contains boost samples, etc. This means that the estimates from the annual LFS dataset have a smaller standard error. However, the annual datasets are not available in time to feed into the ASHE estimation procedure. Consequently, the dataset for the spring quarter, which corresponds to the ASHE survey data, is used since it is available about six weeks after the end of the quarter.

Weights

For the main part of the sample obtained from National Insurance number (NINo) records the weight is the product of a design weight based upon the stratification at the time of selection and a calibration weight based upon the poststratification resulting from the survey responses.

Note that sample selection from the PAYE system is not stratified and each individual has an equal chance of being selected. Hence the design weight for all individuals is the same, and is given by 100 times the number of observations on the sample file (about 240,000) divided ▶

▶ by the number of responses (about 160,000), that is, about 150.

For data coming from sources other than NINo the design weight will be determined by the probability of selecting the chosen business. Since the LFS totals cover all employees (including non-PAYE), the calibration factors are determined in the same way as for the main sample.

Comments on the weighted results

For estimates of pay for previous years released in an article on the National Statistics website on 15 October, the effect of weighting is that:

- different results can be obtained from using different LFS totals for calibration; however, these would be relatively small; and
- weighted estimates are higher than unweighted ones.

The higher estimates generated by weighting may seem counter-intuitive, since the main exclusion from the NES was those individuals outside the PAYE system. However, poorer response rates for employees in high paying occupations more than offset the bias from the PAYE exclusion. In other words, higher-earning employees had been underrepresented in the unweighted sample, and weighting corrects for this. A full investigation was undertaken to determine the contribution each individual made to the difference in estimates when weighting was applied to confirm the nature of the impact that weighting brings.

Standard errors for weighted quantiles

ONS has developed methodology to produce unweighted and weighted estimates of the standard errors of levels of quantiles and the differences in them. In the case of the former, this is by using formulae, and for the latter, the bootstrap method. The development of standard errors for estimates is an important factor in allowing ONS to revisit the criteria against which estimates are judged fit for publication on grounds of quality; this issue is considered further below.

Estimates of the weighted median, upper and lower deciles and upper and lower quartiles, and their standard errors, have been calculated. The weighted estimates and their standard errors are greater than the unweighted ones – behaviour that has been seen already in the estimates of the mean.

The results of analyses of standard errors also reflect the skewness of the distribution of weekly pay. The standard error of quantiles increases with the quantile (for example, the standard error of the tenth percentile is smaller than that of the 25th percentile, etc.). The large and outlying pay records, and the relative sparsity of them, make reliable estimation of the upper quantiles more difficult (that is, the standard errors will be larger) than the lower ones, where there is a greater density of similar values.

The standard errors of a number of the quantile estimates are lower than one might expect, but there is a reason for this: it occurs when a weekly wage is roughly equivalent to an annual salary that is a 'round number', for example £28,000. There tends to be a propensity for employees to be paid in such round number salaries and this causes a bunching in the distribution of pay. The standard error of a quantile estimator taking the value of such a salary, or its weekly equivalent or nearby, will therefore be smaller than if the quantile estimate had happened to be a non-round number annual salary equivalent. The same effect can be seen for weekly pay that equates to a round hourly rate, and in other similar ways.

Sample undercoverage, supplementary surveys and imputation

As noted in the previous section, the target population for the ASHE is all employees. However, employees in businesses that are not included on the interdepartmental business register (IDBR), which is based on information from both PAYE and VAT registrations, cannot be identified and so are excluded from the survey. Businesses of this type are typically organisations where the turnover of the business is below the VAT threshold and/or where the employees earn less than the PAYE threshold. This means that the ASHE-based data on earnings are always likely to overestimate average levels of pay, and potentially could miss an important group of employees at the bottom of the pay distribution. However, the extent of this specific bias is thought to be small since the total number of businesses in this area of the economy is estimated at 1.8 million enterprises encompassing an employment (proprietors and employee) total of 0.9 million. The employee component is thought to be very small, although estimates are not available.

Even within the framework of the IDBR, the sampling frame based on the PAYE system is still inadequate to allow ONS to describe the total population of employees, since it excludes the majority of those employees who do not appear in the PAYE system. Thus, ONS looked at how it might be extended to include businesses with employees but

without PAYE systems, and businesses with employees outside of their PAYE systems. The former are termed 'VAT-only' businesses in the context of the ASHE. These supplementary samples are added to data obtained in the main ASHE and weighted to the LFS population of employees. This reduces the impact of the non-sample bias.

VAT-only businesses

An employee in a VAT-only business will, by definition, earn too little to appear in the PAYE system. That is not to say they are poorly paid, for example an employee paid £10 per hour but working very few hours might not earn enough to merit paying income tax and so not be included in PAYE. The VAT-only sample has different properties to the IR PAYE sample in that all employees identified within an enterprise are included in the scope of the supplementary survey (as opposed to just 1 per cent of the IR file). To obtain data for these employees a selected business is first sent a questionnaire that asks if they have any employees paid outside of the PAYE system. If that is the case then an appropriate number of questionnaires are sent to the business so that they might provide the survey data needed. The 2004 ASHE includes data from a random sample of 5,100 businesses.

Off-PAYE employees

The second area where the current sample underenumerates individuals is in businesses with a PAYE registered payroll system that employ staff that are paid from outside of this payroll. Employees of this type might be loosely or casually attached to the enterprise and should earn below the PAYE threshold. To assess the feasibility of collecting data from this subset of the population a small survey of local units in the hotels and restaurants sector was undertaken. The survey showed that it is very difficult to obtain data for employees of this type. Primarily, this is because the identification of the employees within the relevant businesses is time-consuming and the willingness of businesses to discuss the pay arrangements for employees of this type is low. Consequently, ONS concluded that conducting a supplementary survey of these units in 2004 would not be practical. It is, however, important to note that while the employees will be excluded from the sample set they will be included in the population weights obtained from the LFS.

Non-response

The final source of bias that is addressed in the new design is attributable to non-response. This takes two forms, unit non-response and 'exemption'. The latter is due to employees changing their job between sample selection and the survey reference date, or because the PAYE system fails to reflect job changes at the time that the sample is selected. This is a significant issue in respect of the NES, with around 12 per cent of the NES 2003 sample responses suggesting that the employees selected from the PAYE system had left the employment indicated by the IR's system. To address this issue the 2004 ASHE has included a second despatch of questionnaires where an employee was said to have moved jobs. For these cases the employees' details were matched to a subsequent extract from the PAYE system and the new employer identified. The new employer was then sent a questionnaire and data for the employees sought. This supplement to the survey identified around 1,384 employees and elicited a 73 per cent response. Of the 1,006 questionnaires returned, 52 per cent provided data for the employee, showing that conducting this supplement to the main survey can produce an important gain in sample size.

The issue of non-response has also been assessed as part of the design of the ASHE. In this case the ONS identified a sample of approximately 4,500 employees from within the ASHE sample for whom no response had been received eight weeks after the required response date. This sample was then subject to an intensive response-chasing exercise primarily to identify whether the non-response was in some way non-random. If this were the case it would be possible to use the data to adjust for non-response in a better way than through simple weighting. However, because the follow-up survey is undertaken at the end of the survey processing cycle the results will not be available for use until the 2005 results are processed.

Imputation

While the foregoing sets out how ONS will handle unit non-response in the ASHE, a different approach has been developed to deal with 'item non-response'. This is another area that affected the NES in the past, and while the issue was not a significant problem for processing when the survey results were published in an unweighted form, it is more problematic with the weighting methodology underpinning the ASHE. This is because item non-response, where a questionnaire is returned by a respondent but in an incomplete form, would require the derivation of different weights for different ▶

variables in the survey. While this is technically feasible, it is time-consuming. To address this issue the survey will adopt imputation for those responses where the form is incomplete.

The stochastic imputation method uses a 'donor' approach, where responses from individuals with similar characteristics to the employees with the missing information are used to donate an estimate of the missing variable, forming 'imputation classes'. The variables that will be imputed for when missing are:

- overtime hours;
- overtime pay;
- annual pay;
- normal basic hours; and
- residual weekly pay.

The choice of imputation classes is based partly on the results of the analyses completed to determine optimal stratification supplemented with variables that are relevant to pay. The resulting imputation classes are determined by the following variables:

- two-digit standard occupation class;
- region, where region one was classified as London and the South East and region two as the rest of the country;
- sex;
- adult rate marker; and
- age group, where it takes three values depending on whether the respondent is aged less than 18; between 18 and 21; and greater than or equal to 22.

In developing the imputation method, ONS compared imputed estimates with true values to assess how well the imputation process preserves true values. The analysis showed that true values are well preserved, with no significant difference between the distributions obtained using the true values and imputed values.

Questionnaire redesign, release criteria and the output dataset

A further phase in the development of the survey concerns the redesign of the survey questionnaire. The questionnaire for the 2004 survey was printed on two sides of A4 paper and despatched with a single set of guidance notes to contributors. Thus, all businesses received just one set of guidance notes even if they were required to complete separate questionnaires for a large number of employees. The methodology review of the survey concluded that this version of the questionnaire was substandard and in need of change to allow ONS to capture data accurately, especially in respect of the pay and hours data used to derive an hourly rate of pay for employees. To address these issues ONS's Data Collection Methodology (DCM) Unit undertook a programme of work to review the user requirement, assess emerging user needs in the context of the survey (for example, pensions issues) and design a new format for the questionnaire. This new format was then taken through a programme of cognitive testing with businesses of all sizes and in all sectors of the economy. The new design, which conforms to theoretical best practice, is nearing its final form.

The final design will be informed by the outcome of an analysis of a field test conducted in parallel with the 2004 ASHE. This field test involved ONS selecting a random sample of 5,000 employees that was extracted from the main ASHE sample. The aim of the field test is to allow ONS to assess whether the reworded questions included in the new questionnaire can be answered readily by businesses. A second objective for the test is to indicate whether the inclusion of guidance notes as part of the questionnaire, rather than as a separate set of instructions, reduces item non-response and incorrect responses, and improves accuracy in respect of the target variable. The field test should also show whether a switch to a longer questionnaire affects response rates adversely: the version tested was printed on six sides of paper rather than the two that users are familiar with. This approach encapsulates the greater methodological rigour that ONS is bringing to the design of its survey instruments and allows ONS to report with greater confidence the results of its surveys.

The redesign of the questionnaire and the cognitive testing exercise allowed ONS to assess the quality of data that were obtained within the NES on bonuses. This is a problematic area, and one that impedes the capacity to make like-for-like comparisons with the Average Earnings Index (AEI). Following the review of the questionnaire, ONS is likely to stop asking employers to provide data on bonuses that are paid outside of the reference period but which relate to work undertaken in the reference period. Data of this type are only available in a real sense from businesses paying bonuses in May or June and, to a limited extent, in July. This reality reflects the response deadlines for the ASHE, where the survey data are provided in respect of April and the survey take-on and validation ends in August. Thus the majority of annual bonuses, which the survey used to compile the AEI shows are paid in December, January

and March, are missed by the current questionnaire design. Instead, ONS will capture data on total bonuses paid in April, and so allow the generation of a figure on a comparable basis to the AEI for that month. Additionally, employers will be asked to indicate the part of these bonuses that relates to work undertaken in April. Supplementing these questions will be information on bonuses paid in the tax year, as a component of total annual pay. Dividing this annual total by 12 to get an average monthly level of bonuses will give some indication of the impact of irregular bonuses on a month by allowing a comparison with the bonus data that relates to work undertaken in April.

These changes should improve the quality of the estimates from the 2005 survey compared with those obtained using the existing survey questionnaire.

The output dataset

This final discussion looks at the outcome of work to revise the release criteria for estimates produced in the ASHE, which ONS will apply to the weighted results from the new survey in 2004. As a result of this work, ONS will reduce the number of standard tables that are produced each year, and replace them with a shorter summary set of outputs that better meet the immediate needs of users for important indicators on earnings statistics.

NES release criteria

Historically, NES data were assessed for their quality according to the following criteria.

- For NES published tables: if the sample size was 30 or more and the relative standard error of the mean estimate of pay/hours was less than 5 per cent, then the estimate would be published.
- For ad hoc queries and on NOMIS®: if the sample size was 5 or more an estimate would be given. If the sample size was less than 30 or the relative standard error of the mean estimate of pay/hours was more than 5 per cent, then it would be indicated that the estimate was of poor quality and that such estimates should not be used in publications.

These criteria were applied to all statistics (not just means) including proportions and quantiles (for example, in assessing the quality of the median, the standard error of the mean was examined). The second set of criteria allows for very detailed estimates to be produced, many of which will be of poor quality. Anecdotal evidence suggests that the release of estimates under these criteria has led to misuse or at least stretched use of the data. These arrangements have been in force for many years and their provenance is uncertain. It is likely, though, that the basis of the second of the criteria was motivated by demand for very detailed NES estimates.

ASHE publication

ONS will significantly change the way the annual earnings data are presented. The NES data were issued in a National Statistics First Release, accompanied by a more detailed set of tables available on-line at the National Statistics website. The on-line tables are an electronic version of the paper publication that has historically been produced for the survey. Following the release of the summary volume, the NES results set was then issued, again on-line, in a further seven volumes of data tables presenting results by region, industry, occupation, collective agreements, etc. The content of the volumes has not been revised to any great extent since the survey's inception, and as a result the information presented often confounds rather than informs. The new publication for ASHE will amend the presentation of the results, such that headline statistics for various subgroups of the population will be available in a single volume, with all other requirements being met on request. The focus of the results will switch from estimates of mean pay to those of median earnings.

The disclosure rules for ASHE have also been changed to bring them in line with wider ONS practice. Estimates for a table cell with less than three responses are considered as potentially disclosive, and so are suppressed. In addition, cells are suppressed if they fail the ONS rules for dominance. The dominance rule determines whether a cell of a table is disclosive owing to a small number of respondents contributing to a large proportion of the total. This allows for the publication of more estimates under ASHE than was allowed under NES.

In addition, quality measures in the form of coefficients of variation (CV) will be published for all ASHE variables. The coefficient of variation is the standard error of an estimate divided by the estimate. To help the user in interpreting the quality of estimates presented in tables a new quality key has been introduced (see p464).

Estimates are marked in different colours according to their CV value in relation to quality thresholds. For example estimates with a CV of greater than 10 per cent but less than or equal to 20 per cent are marked as 'acceptable'; such estimates should be used with caution. Estimates ▶

Key
Precise CV <=5 per cent
Reasonably precise CV >=5 per cent and <=10 per cent
Acceptable CV >=10 per cent and <=20 per cent
x = unreliable CV >=20 per cent or unavailable
.. = disclosive n/a = not applicable

where the CV is more than 20 per cent are suppressed as they are considered to be unreliable.

A further improvement over NES is that the new ASHE publications include responses for Northern Ireland with the first results, rather than these being added later as with NES.

Conclusion

The methodology underpinning ASHE 2004 will:

- introduce the weighting of results to the population of jobs measured by the LFS, imputation for item non-response and sample error estimation;
- extend the coverage to include employees in VAT-only units held on the IDBR, and people who change or start a job between sample selection and the survey reference period;
- redesign the survey questionnaire, planned to be introduced for the 2005 survey;
- amend the results release criteria; and
- change the nature of the survey results publication.

Publication of results

Results and back series for 1998 to 2003 using the same imputation and weighting methods as defined in the ASHE methodology but applied to the NES data sets were published on 15 October 2004. The results were published on the National Statistics website using the new publication layout and quality criteria. At the same time, two articles were published. The first describes the impact of applying the new methodology to the 1998 to 2003 NES data. The second describes changes to the methodology used to compile estimates of low pay and the impact these changes make to the estimates for 1998 to 2003. These articles will both appear in forthcoming issues of *Labour Market Trends*.

Results for the April 2004 ASHE survey and revised results for the 2003 survey were released on 28 October 2004. These use the new ASHE methodology, and for the 2004 survey results were published both including and excluding responses from the supplementary surveys so that comparisons can be made with earlier results. The results were published on the National Statistics website using the new publication layout and quality criteria. Results and back series for 1992 to 1997 will be released as soon as they have been quality assured.

Further information

For further information, contact:
Chris Daffin,
Room D101,
Office for National Statistics,
Cardiff Road,
Newport NP10 8XG,
E-mail: earnings@ons.gov.uk,
Tel. 01633 819023.