

QA of work location(wpost) problem

Issue/area of concern

The Annual Survey of Hours and Earnings (ASHE) records an employee's home and workplace location. This is of great potential value in assessing the role of geographical location in labour market analysis and in estimating geographical impacts of policy. There is concern about the accuracy of the work location variable as the underlying survey question is pre-filled by Office for National Statistics (ONS) with the registered Pay-As-You-Earn (PAYE) address of an employer. Employers are required to check and correct the address if inaccurate, creating the potential for measurement error if this checking and correcting is non-random.

Findings

We tested the hypothesis that there is measurement error in the ASHE work location variable, leading to possible errors in employee work location for multi-site organisations. This is likely to be a problem for multi-site companies who must complete forms for each individual. Single-site businesses and employers who supplied payroll records electronically to ONS in bulk (rather than completing a survey form) provided comparison groups.

Comparing ASHE work locations to workplace locations recorded for the employer in the Business Structure Database (BSD), multi-site enterprises completing the paper survey showed a higher proportion of ASHE employees working at the head office location than organisations with a single site and/or completing electronic responses. Analysis of commuting distances also supported the hypothesis: we find considerably longer commuting distances for employees working at enterprises returning survey forms, as compared with those making bulk submissions electronically. These differences are also positively associated with the number of sites, private sector employers, and certain industries.

Changes made to data

A number of new variables have been created for this analysis and are available for future use by researchers. They are based on Census Output Areas (the smallest geographical unit in both ASHE and the BSD), including: total number of employees by enterprise (num_ent_emp), head office marker (head_office), number and proportions employed at head office (HO_emp and HO_prop_emp), and straight-line distance from employee home (hpost) to work location (wpost), using easting and northing point coordinates (diff_en). However, at present, there is no information on whether the workplace location has been checked.

Recommendations to researchers

There are various short-term measures one can use to reduce the potential for unrepresentative descriptive statistics or model estimates arising because of workplace measurement errors. These include creating sub-samples and checking whether results are robust to including two new control variables in regressions - a dummy for whether the survey questionnaire was sent to an employer and the number of local units (plants/workplaces) in the enterprise. It is likely that descriptive statistics are more sensitive to workplace location errors than regression analysis.

The WED team is investigating whether it is possible to identify whether employers have edited the 'work location' field; this would enable some additional measure of quality assessment.

Details of QA

Analysts: Damian Whittard, Felix Ritchie

QA Code file: HQ analysis, P:\Working\code\qa_03_HQ-analysis_v06.do

Variable creation code: can be found in the main ASHE creation code

This version: November 2021

Detailed analysis

Summary:

- The variable 'wpost' records an employee's work location. There is concern over its accuracy.
- By combining the ASHE and BSD datasets, we can test whether this variable is likely to suffer from non-random measurement error.
 - Initial tests compare head office employment of the same enterprises as recorded in the two different datasets.
 - Subsequent tests compare the commuting distances of employees working for enterprises making paper submissions compared with employees working for enterprises making electronic submissions¹.
- The study suggests the existence of non-random measurement error in the recording of the workplace location in ASHE among organisations where the survey is administered via a paper questionnaire. Regression analysis reveals that a large proportion of the variance in (implied) measurement error is accounted for by:
 - private sector employers, particularly from the finance/law, utilities and construction sectors.
 - multisite employers with more than 10 sites.
 - region of registered workplace

The study recommends a number of alternative approaches to improve the quality of the data and reduce the potential for biased estimates. Full details are available in the working paper from the WED website (Whittard et al., 2021)².

Problem(s):

- In late 2019, during a discussion on potential issues in relation to the ASHE survey, the Northern Ireland Statistics and Research Agency (NISRA) suggested there was a possibility of errors in the collection of the locations of an employee's workplace (and home) location employment in ASHE.
- To test this hypothesis, we:
 - (i) linked the BSD and ASHE dataset at the level of local unit to compare head office employment;
 - (ii) compared reported commuting distances between employees working for enterprises making paper submissions with those making electronic submissions, and between single-site and multi-site enterprises.
- The difficulty with (i) is that there is no direct method to link the two datasets. This is because:
 - a) the local unit reference in the ASHE is not able to be linked directly with the BSD LUREFs;
 - b) although full postcode data is available in ASHE, in the BSD it is truncated and therefore this 'proxy' local unit marker cannot be linked.
- The solution was to create a geographical information marker at the level of the census output area, (COA) for both the enterprise (proxy for the headquarters) and the local unit (workplace address), which was consistent between the ASHE and the BSD.
- The difficulty with (ii) is that a marker for those making electronic submissions was only included in the ASHE dataset from 2016. This limits the analysis to recent years.

¹ By 'electronic submissions' we mean those employers providing data to ONS directly from their payroll records, i.e., a bulk transfer of data. There is an option for employers to complete the ASHE survey form electronically, but this needs to be done on an employee-by-employee basis, so we also treat these as 'paper' submissions.

² Whittard, D., Ritchie, F., Phan, V. Forth, J. Stokes, L. Bryson, A. Singleton, C. and McKenzie, A. (2021) Exploring the workplace location problem in the Annual Survey of Hours and Earnings. WED Working Paper

- The analysis suggested the data was subject to non-random measurement error, which has important implications from a policy perspective. For example, it has the potential to affect estimates of average regional earnings differences, with their attendant implications for the levelling-up agenda.
- As such, approaches to improve the quality of the data and reduce the potential for biased estimates have been suggested.

Strategy:

- Focus on the ASHE datasets for 2016, 2017 and 2018.
- Merge the ASHE annual datasets with COA data for the enterprise, including the head office (a Head Office lookup file is now created as part of the core dataset construction)
- Create a number of markers on ASHE including:
 - head office, multisite employers, special arrangement³ ('electronic') returns, 1-digit Standard Industrial Classification (SIC) codes
- On an annual basis and for the three pooled years, calculate the number and proportion of ASHE respondents observed working in the head office by enterprise
- Compare the BSD and ASHE proportions of employees working at head office by the total number of employees in an enterprise
- Calculate the average straight-line distance between an employee's home address and the enterprise address for those making paper submissions, compared to those making electronic submissions
- Run multivariate regressions to test the importance of a company making electronic submissions on the accuracy of the response, while isolating the relationship by controlling for all other factors.

Findings

Table 2 in the main study compares head office employment for the same enterprises using the data recorded in the BSD (columns 1 and 2) and comparing that with data in ASHE (columns 3 and 4). This means that columns 1 (BSD) and 3 (ASHE); as well as columns 2 (BSD) and 4 (ASHE), provide information on the same businesses recorded in the two different datasets. Theoretically the number of observations in columns 1 and 3, and 2 and 4, should be equal. However for processing purposes (see Whittard et al., 2021 for full detail) a handful of BSD observations had to be dropped from the sample (approximately 50).

In terms of multi-site companies, columns 2 and 4, shows that the same companies in ASHE report a considerably larger proportion of employees recorded as working at their head office (80%) than is recorded in the BSD (61%).

³ Approximately 350 of the largest employers have 'special arrangements' in place. This allows them to make to make a bulk electronic submission generated from their payroll system. A benefit of this is that the work place location is more likely to be correct, as it is directly entered from the employers system and not pre-filled by ONS.

Table 2: Numbers of enterprises and proportions of their employees based in head offices, by single site

and multi-site enterprises, ASHE vs BSD (2018)

		BSD (1)	BSD restricted sample (2)	ASHE (3)	ASHE restricted sample (4)
single	Number of obs.	40,090	32,778	40,142	32,816
	Mean	100%	100%	81%	99%
	Median	100%	100%	100%	100%
multi	Number of obs.	16,078	11,172	16,080	11,174
	Mean	54%	61%	56%	80%
	Median	55%	66%	67%	100%
unmatched	Number of obs.	-	-	976	-
	Mean				
	Median				

Source: ONS, BSD and ASHE

Table 4 in the main study shows employees working for multi-site companies making paper submissions are reported as living considerably further away (28km) than employees either working for single site enterprises or multi-site enterprises that make electronic submissions (<17km). This is consistent with the hypothesis that there is potential non-random measurement error in the workplace location variable for those enterprises making paper submissions.

Table 4: Distance (in kilometres) between recorded employee work and home addresses in the ASHE (2018)

Organisation type	Observations (1)	Average (2)	10th percentile (3)	50th percentile (4)	90th percentile (5)
Single site paper	56,647	16.6	0.5	5.7	29.5
Multi-site, paper	96,633	28.2	1.1	7.7	54.4
Single site, electronic	233	8.5	0.5	4.9	19.5
Multi-site, electronic	25,200	16.4	0.9	5.5	30.9

Source: ONS, ASHE

Multivariate regression analyses (tables 8 and 10 of the main study) further confirm the hypothesis of potential non-random measurement error in the workplace location. To demonstrate this, an extract of table 8 and 10 have been included here – the extract just includes the results for the ‘special arrangements’ variable. The results demonstrate that controlling for all other factors, enterprises making electronic submissions (noted as having ‘special arrangements’) negatively affects the probability of having employees as recorded as working at the head office, while it also considerably reduces the distance (by 9km) that an employee is recorded as living from work.

Table 8: Regressions on the proportion at head office, all multisite enterprises and multisite enterprises ignoring ‘shell’ companies (2018)

Dependent variable: proportion at head office	OLS (1)	OLS - restricted (2)	Tobit - restricted (3)
Special arrangements	-0.14***	-0.11***	-0.14***

Table 10: Distance travelled to work, all employees and adjusted for enterprise clustering (2018)

Dependent variable: distance travelled to work	Distance - all observations		Distance- multisite only	
	Kilometres (1)	Logs (2)	Kilometres (3)	Logs (4)
Special arrangements	-9.59***	-0.16**	-9.50***	-0.16**

The results from the descriptive statistics and regressions provide consistent support to the hypothesis that employees working for employers who make an ASHE paper submission are more likely to have their work location incorrectly recorded. This is important from a policy perspective as it can affect analysis of commuting patterns as well as regional wage disparities.

Does it matter?

Yes. Table 11 gives a flavour of the potential effect on regional wage differentials by recording the percentage reduction in the mean gross earnings between the full sample and a restricted ‘high certainty’ sample. The high certainty sample excludes those returns which we believe may suffer from non-random measurement error – i.e. multi-site companies making paper submissions.

Table 11: Comparison of median annual gross wage between full sample and excluding multi-site companies making paper submissions.

Government Office Regions	Percentage reduction in mean gross earnings having removed multi-site companies making paper submissions
North East	16%
North West	12%
Yorkshire & Humberside	10%
East Midlands	7%
West Midlands	11%
South West	10%
East	7%
London	3%
South East	10%
Wales	15%
Scotland	7%

Source: ONS, ASHE

Table 11 shows that all regions record a fall in the average wage rate, but the fall is lowest in London and the greatest is in the North East. This suggests that using the high certainty sample would increase the regional pay gap.

The authors recognise that the approach used here is blunt and, as such, we highlight that these results should only be seen as illustrative. This is because cases that were removed to create a ‘high certainty’ sample may be unusual for many reasons, and not just because they potentially suffer from non-random measurement error.

At the time of writing, the research team are in discussions with ONS about the potential to identify which companies have amended the work location variable. If this data becomes available, then there will be an opportunity to develop a more nuanced approach to estimating regional pay differentials. This will be based on correcting potential measurement error using probabilistic modelling.

Limitations

The authors acknowledge that the findings in this paper are provisional, indicative and do not provide conclusive evidence of non-random measurement error. However, given the potential importance of the

findings, we are keen to publish our results to generate feedback and discussions, while recognising this study is still a work in progress.

The main limitation of this research is that we are only able to look for potential evidence of non-random measurement error, without actually being able to observe it. This makes correcting for any potential measurement error challenging, albeit some suggestions were presented above. In order to further develop the quality of the data and analysis, we plan to undertake the following:

- Develop a probabilistic model which can be used to weight the data. This can be used to improve the quality of the data and any subsequent analysis.
- Develop the econometric analysis to include interaction terms. This has the potential to expand the understanding of the relationships among the variables in the model and allow additional hypotheses to be tested

Conclusion

The results suggest that there is potential non-random measurement error in the workplace variable of employees working for enterprises making paper ASHE submissions.

In the short term, this study suggests that improvements to the data and model estimates can be achieved by creating high quality sub samples and including additional control variables, such as a paper questionnaire dummy variable and a variable recording the number of local units.

In the medium term, however, it suggests that improvements in the data collection process should be made – or that, at least, information should be made available to researchers on whether defaults were changed.