

Feasibility study exploring the potential to link administrative education data to the Annual Survey of Hours and Earnings

Dr Claire Crawford

UCL Centre for Education Policy and Equalising Opportunities

March 2022



Summary

This feasibility study is part of the Wage and Employment Dynamics (WED) project, which is seeking to provide new insight into the dynamics of earnings and employment in the UK, by enhancing the Annual Survey of Hours and Earnings (ASHE) through linkage to other survey and administrative data.

This feasibility study focused on the potential for linking education data. Specifically, it set out to:

- 1) Scope the possibilities for linking administrative education data to ASHE and, as far as possible, get a sense of any legal and practical barriers to pushing those linkages forward.
- 2) Summarise the types of research questions that could potentially be addressed by each linkage and try to provide some sense of which research questions are most pressing from a policy perspective.
- 3) Scope, as far as possible, any existing plans for other linkages which may deliver answers to some or all of these questions, and hence where any ASHE linkages could add most value.
- 4) Provide recommendations for which linkages should be taken forward and in what order.

As the study has been funded by ADR England (Administrative Data Research England), its focus is primarily on the feasibility of linking education data for individuals living in England, although we touch briefly on the possibilities in other UK nations.

In England, we considered the pros and cons of linking data from the Higher Education Statistics Agency (HESA) on participation in higher education; the Individual Learner Records (ILR) on participation in further education; and the Longitudinal Education Outcomes (LEO) data, which combines ILR and HESA data with school records from the National Pupil Database (as well as tax and benefit records).

Our review suggests that there is significant appetite within the Department for Education (DfE) (primary data controller of LEO data for England) for a link between ASHE and LEO. (Linking in ASHE would provide additional information, including on hours of work and occupation, for a subset of individuals in LEO, providing benefits for that dataset as well as for ASHE.) From a research perspective this encompasses the same advantages as a link to standalone ILR data and has advantages relative to linking in standalone HESA data (especially in the short-term). We would therefore strongly recommend that ADR UK (or researchers funded through ADR UK) work closely with DfE officials (and officials from departments contributing data to LEO in England – namely, Department for Work and Pensions (DWP), Her Majesty's Revenue and Customs (HMRC) and HESA) to ensure this linkage occurs, and that the resultant linked data is well documented and can be shared readily with external researchers. If this option does not prove feasible, then there would be merit in pursuing a link to standalone HESA data instead.

There also appears to be strong merit to considering links between ASHE and administrative education data in all UK nations. At a minimum, HESA data is a common resource that could be linked and coded in the same way, facilitating better answers to a number of interesting research questions. However, data from different nations is currently made available to researchers in different places, preventing the benefits of a UK-wide spine of wage and employment data linked to education data from being fully realised at the present time.

Suggested next steps

- Maintain a dialogue with officials at DfE and the devolved administrations regarding plans for linking ASHE and LEO, and identify how support from ADR UK could be used to support and develop these plans.
- Keep in touch with representatives from all UK nations regarding current and future plans for linking ASHE to other data, to share learning and identify opportunities for data development.
- Join or instigate discussions regarding the possibility of sharing data – including, but not limited to, linked ASHE data – across the UK, to maximise the benefits of these linked data projects.

1. Introduction

The Wage and Employment Dynamics (WED) project¹, funded by ADR England², set out to provide new insight into the dynamics of earnings and employment in the UK, by creating a sustainable, documented wage and employment data ‘spine’, built around the Annual Survey of Hours and Earnings (ASHE).

ASHE is a longitudinal survey which collects information about the employment of a 1% sample of employees in the UK (comprising roughly 300,000 individuals). The data is collected directly from employers and contains detailed information about gross pay (including overtime and incentive pay), hours and occupation, as well as some basic background characteristics, such as gender.

ASHE is “the most comprehensive source of information on the structure and distribution of earnings in the UK”³, but is limited in a number of respects. Firstly, it only includes an annual snapshot, collected in April each year, limiting the possibility of looking at the dynamics of employment and earnings in between these periods. Secondly, it contains very little information about the individuals other than their employment details, meaning that the scope for understanding the factors explaining or contributing to observed dynamics are extremely limited.

The WED team are trying to supplement the information available in ASHE to address these limitations, by linking in data from other survey and administrative datasets.⁴ The team has already linked in data from the 2011 Census for individuals in ASHE in England and Wales, and has secured permission to link to Real-Time Information (RTI) data from HMRC.⁵ The link to RTI data will offer insight into individuals’ employment status in between ASHE snapshots, as it provides start and end dates of employment spells. It will also include information from self-assessment records, meaning it will be possible to observe individuals moving in between employment and self-employment.

The link to the 2011 Census provides more background information on individuals and their households, including ethnicity, health status and – of particular relevance for this study – highest educational qualification. This will provide important new insight into the dynamics of employment and earnings gaps by ethnicity, for example, and allow more detailed exploration of the characteristics and family circumstances of individuals in low paid employment than is possible in most survey datasets.

The link to information on education is particularly important, as it opens up the possibility of identifying the returns to different educational qualifications using more detailed and potentially

¹ <https://www.wagedynamics.com/>

² <https://www.adruk.org/our-work/browse-all-projects/wage-and-employment-dynamics-in-britain-143/>

³ www.ons.gov.uk/surveys/informationforbusinesses/businesssurveys/annualsurveyofhoursandearningsashe

⁴ See the [Wage and Employment Dynamics website](#) for further details of the team’s plans.

⁵ The geographic scope of this linkage has not yet been formalised.

more accurate information than is available in most other data.⁶ It also allows greater understanding of the way individuals move into and progress through different occupations, and how this may vary across different groups.

However, the information on education available from the 2011 Census is very limited. It only provides information on an individual's highest educational qualification, and this information is categorised using very broad groups. For example, including all those with undergraduate and postgraduate degrees in all subjects and from all institutions together, whereas we know that there are significant differences in returns to degrees at different levels, in different subjects, undertaken at different institutions and awarded different degree classes.⁷

This feasibility study explores the potential for linking in other data on education participation and attainment to overcome these limitations. Specifically, it set out to:

- Scope the possibilities for linking administrative education data to ASHE and, as far as possible, get a sense of any legal and practical barriers (including department appetite, resource constraints, etc.) to pushing those linkages forward.
- Summarise the types of research questions that could potentially be addressed by each linkage and try to provide some sense of which research questions are most pressing from a policy perspective.
- Scope, as far as possible, any existing plans for other linkages which may deliver answers to some or all of these questions, and hence where any ASHE linkages could add most value.
- Provide recommendations for which linkages should be taken forward and in what order.

It was conducted using informal interviews with a mixture of data owners and expert users of the data (listed in the acknowledgements section).

As the study is being funded by ADR England, its focus is primarily on the feasibility of linking education data for individuals living in England. However, significant interest has also been expressed in the feasibility of creating similar linked resources in other countries of the UK, thus we additionally consider the appetite for and feasibility of undertaking similar exercises in Northern Ireland, Scotland and Wales.

The report proceeds as follows:

- Section 2 summarises the primary education datasets considered for linkage to ASHE for individuals in England and reviews our understanding of the pros and cons of each linkage, as well as the appetite for and feasibility of linking ASHE to each of these datasets.
- Section 3 provides a briefer overview of the same questions for Northern Ireland, Scotland and Wales.
- Section 4 makes recommendations regarding priorities for linkage and summarises some suggested next steps.

⁶ Estimates of the returns to education in the UK – England in particular – have, in the past, been undertaken using the Labour Force Survey (e.g. Walker and Zhu, 2011), which relies on self-reported information from workers on hours and pay, which researchers must use to estimate hourly wages, or using the Longitudinal Education Outcomes (LEO) data (e.g. Britton et al., 2021a,b) which provides information on annual earnings from HMRC, but does not contain information on hours, so hourly wages cannot be calculated.

⁷ See, for example, Britton et al. (2020a,b, 2021b); Naylor et al. (2016).

2. Administrative education data linkage in England

Datasets considered

The starting point for the project was consideration of a link to **Higher Education Statistics Agency (HESA)** data.⁸ HESA data collates annual records on both students and staff at higher education (HE) institutions in the UK. Given that sample sizes of staff working at higher education institutions who are also in ASHE are likely to be too small for meaningful data analysis, our focus is on linking student HESA records.⁹

Student HESA records are available for all individuals undertaking HE at any UK institution since the academic year 1994-95. They act as a census of those registered to attend in the first term of each academic year, providing information about:

- the course students are registered to attend (such as level, subject and mode, full-time or part-time)
- some demographic characteristics – including gender, ethnicity, age and measures of social background (such as self-reported parental education for younger entrants)
- measures of prior attainment (in the form of an A-level and equivalent tariff points score)
- an individual's circumstances whilst at university, such as whether they live at home or in university-owned accommodation
- the continuation status or outcome of an individual's engagement with HE: for example, whether students left the course prior to completion (and the reasons for this), as well as their degree class (where relevant).

A link between ASHE and HESA data would add significant value to the education qualification information already matched in from the 2011 Census, addressing at least two important research questions. Firstly, it would provide information on HE subject, institution, and degree class, which – as highlighted above – have been shown to significantly affect graduate earnings, and hence the returns to education. Secondly, it would enable exploration of student work patterns, as well as the employment, occupation and earnings dynamics of individuals with different experiences of student employment.

A disadvantage of a link to HESA data alone, however, is that it would only provide supplementary information for those who have attended university, missing the roughly half of the population of 17-30 year olds (and higher proportion of the full working age population) who do not participate in HE.¹⁰

One option would be to supplement a link between ASHE and HESA with an additional link designed to capture information on further education (FE) participation and attainment. In England, such data is available from **Individual Learner Records (ILR)**, for all those engaging in FE since 2002-03, providing information on the type, subject and level of qualification undertaken, including

⁸ <https://www.hesa.ac.uk/data-and-analysis/students>

⁹ HESA also collects information on outcomes six months (and, for a subset of students, 3.5 years) after graduation via its Destination of Leavers from Higher Education (DLHE) survey (now known as the Graduate Outcomes survey). Conversations that have taken place since this report was initially drafted have revealed interest from HESA in linking this data to ASHE as well. These surveys provide information on the type of work graduates are doing that is not available in ASHE (e.g., whether it relates to their degree), as well as information on the way in which individuals found their jobs, perceptions of their job and the activities of those not in work, which would significantly widen the pool of potential research questions that could be explored relative to matching just to the HESA student record.

¹⁰ The latest estimates of the proportion of 17-30 year olds in England participating in HE can be found here:

<https://explore-education-statistics.service.gov.uk/find-statistics/participation-measures-in-higher-education>.

apprenticeships.¹¹ This would again provide much richer information on educational qualifications compared to that available from the 2011 Census, although it is unlikely to be possible to explore returns to all qualifications individually, because sample sizes would not be large enough in all cases. As with HE, it would also enable exploration of the contemporaneous employment experiences of individuals undertaking vocational training, and the firms who are offering such opportunities, e.g., apprenticeships. Unlike with HESA data, however, there would be no information on prior educational qualifications, making it harder to identify the causal effect of obtaining a particular qualification on labour market outcomes.

The [Longitudinal Education Outcomes \(LEO\)](#) data for England potentially offers the best of both worlds. LEO combines information from ILR and HESA with information from the National Pupil Database (NPD), a census of all pupils attending state schools in England together with the national achievement test results of all pupils in England who take them (including those in private schools). These three education datasets have been matched to HMRC tax records and DWP benefit records, and the resultant dataset has recently been made available for research purposes via the Office for National Statistics (ONS) Secure Research Service (SRS).¹²

Specifically, LEO data for England includes the following individuals:

1. Individuals born since 1985 who are 14+ years of age and appear in the NPD (i.e., who have attended a state school or undertaken a national achievement test in England since 2001-02)
2. People who are older than this who have:
 - a. Participated in HE at an English institution since 1994-95 (or, for English learners only, participated in a Scottish or Welsh institution), or
 - b. Participated in FE in England since 2002-03.

It is clear from this description that LEO includes all individuals in the ILR. There is therefore no benefit – at least from a research perspective – to pursuing a link between ASHE and ILR in favour of one between ASHE and LEO. Compared to matching standalone HESA data – which covers individuals attending HE institutions throughout the UK – it would miss information on those from England attending institutions in Northern Ireland. However, less than around 2,000 students domiciled in England attended an HE institution in Northern Ireland in 2019-20¹³, a figure which is likely to have been lower in previous years, suggesting that LEO data would miss relatively few individuals in ASHE from England that standalone HESA data would capture. LEO data also improves upon the information available from standalone HESA data by providing (for those attending school and/or FE in England) very detailed information on prior education attainment, thus enabling analysis of the consequences of following different education trajectories in terms of occupation or hourly wages, for example.

¹¹ <https://esfahelp.education.gov.uk/hc/en-gb/articles/360017077060-What-is-the-individualised-learner-record-ILR->

¹² <https://www.gov.uk/guidance/apply-to-access-the-longitudinal-education-outcomes-leo-dataset>.

¹³ Authors' calculations based on <https://www.hesa.ac.uk/data-and-analysis/students/where-from>.

Practicalities

LEO data for England: the Department for Education (DfE) is the primary data controller for LEO (although permission to link to ASHE would also need to be secured from the owners of contributing datasets – namely HMRC, DWP and HESA¹⁴). Discussions with officials at DfE have indicated significant appetite for a linkage between ASHE and LEO. We have been told that it is at the top of their list of priorities for the development of LEO, both because it will enable exploration of the education pathways that lead individuals into particular occupations, and because it will provide information on hourly wages (as opposed to annual earnings from HMRC records), which is vital to understanding the returns to education – including gender differences therein – in more detail. (DfE's primary interest in the linkage derives from its benefits for LEO, but the same linkage process would also enable benefits for ASHE, which is the focus of the WED project.)

DfE are currently exploring the best way through which to undertake this linkage. Assuming a solution can be found, there are encouraging signs that agreement could be reached regarding a link between ASHE and LEO data for England.

HESA student records: HESA are the data owners. However, ONS holds HESA data from the early 2000s, and is currently engaged in discussions with HESA about taking ownership of this data, including the possibility of using it to link to ASHE. Sourcing further HESA data (e.g., back to 1994-95) would likely require additional resources and could potentially make discussions regarding data linkages more complicated and protracted, although this was not ruled out. In the absence of additional data, a link between ASHE and the HESA data currently held by ONS would therefore exclude individuals who participated in higher education in the 1990s. (A link to LEO data would include these individuals.)

Other linkages

Our review has identified no other linkages in the pipeline which would supplant the benefits of a link between ASHE and some form of education administrative data, especially in the short-term.

3. Administrative education data linkage in other UK nations

Each nation has its own approach to data linkage and sharing, which will be well understood by ADR UK. The following summary therefore touches on this only as it relates to the specific question of whether and how ASHE data could be linked to the education administrative data available in each nation.

Wales

Welsh administrative education data is very similar to what is available in England: it has its own NPD, its own equivalent to the ILR (the Lifelong Learning Wales Record (LLWR)) and HESA data. These datasets can be linked to each other and indeed to other datasets, such as health records, and accessed via the Secure Anonymised Information Linkage (SAIL) Databank (equivalent to ONS SRS).

A LEO dataset for Wales has been created, which contains broadly equivalent data to LEO data in England. It is not currently available for research purposes in the SAIL Databank, but work is underway to explore the possibility of making it available in SAIL.

The SAIL Databank does not currently hold a de-identified version of ASHE for Welsh participants, although a business case requesting this has been submitted. If permission is granted for SAIL to hold

¹⁴ HESA have handed over ownership of their data currently in the SRS version of LEO to DfE, so separate permission to link to this data may not be required from HESA. If a link were to be made using data other than that already held by ONS, it is unclear whether HESA or DfE would be required to consent. However, conversations with HESA indicate an appetite for linkage, so hopefully this wouldn't be a barrier.

a de-identified version of ASHE, then a link to the above-described education administrative data – or, potentially, to LEO data – could be undertaken, with data owner agreement.

This means that, in principle at least, comparable derived variables could be created for both ASHE linked to education data in England, and for ASHE linked to education data in Wales. However, as things currently stand, it appears that the data would be held and made available separately (English data within ONS SRS; Welsh data within the SAIL Databank), and a single secure site through which a combined English and Welsh dataset could be accessed does not sound likely at this stage. This is therefore an important limitation to the aim of creating a UK-wide version of ASHE with comparable linked data. Even where similar data exists in different nations, the resultant linked data may not be available to researchers in one place for combined analysis.

Northern Ireland

Northern Ireland holds ASHE data on its population and has already linked it to the 2011 Census, a de-identified version of which has recently been made available for research purposes via the ONS SRS and its own secure service, known as the Earnings and Employees Study (EES) 2011.

Work is also ongoing to create a new linked education administrative dataset which would form the basis of a Northern Irish version of LEO. This includes a school census, similar to the NPD in England, but which (unlike in England) does not include any attainment data prior to GCSEs and collects only limited demographic characteristics. Attainment data from academic qualifications (GCSEs, A-levels) has only been collected for data sharing and linkage purposes since 2018-19 (replacing a previous way of collecting exam information). In addition, there is HESA data, but no equivalent to the ILR or LLWR.

The value-added of trying to link administrative education data to ASHE in Northern Ireland is therefore less clear at this stage. The sample size is already small (less than 6,000 records), and the number for whom anything other than HESA data would be available is likely to be extremely small at this stage (although this may change as the NI LEO data linkage project develops). There may be merit in considering a link to standalone HESA data, which would, in principle, enable comparable information on HE participants to be made available across the UK. Again, however, the data would most likely be held separately, although there seems to be greater appetite for data to be shared and made available via the SRS (although whether in a single UK dataset or separate national datasets is unclear).

Scotland

Like Wales, Scotland holds some of its education data within its own secure access service (the National Safe Haven), with included datasets all potentially linkable via a single identifier. The datasets currently available focus on school pupils, containing information similar to the NPD in England and Wales. Data on school leaver qualifications and destinations is expected to be added shortly. HESA data is not currently included. Data similar to ILR or LLWR is understood to be held by the Scottish Funding Council, but again is not currently included, nor are there plans for this data to be included in the foreseeable future. There is also a Scottish version of LEO, but, again, this is not currently available in the National Safe Haven.

ASHE data is held for research and statistical purposes but is not currently available via the National Safe Haven. To date, it has not been possible to ascertain whether a de-identified version of ASHE is held, meaning it may be more challenging to implement a link between ASHE and administrative education data in Scotland than in other UK nations, at least at the moment.

4. Recommendations and next steps

England

Our recommendation for England would be to pursue a linkage between ASHE and LEO in the first instance. While it would exclude information for individuals from England participating in higher education in Northern Ireland, the research questions that could be addressed would be considerably enhanced by including information from NPD and ILR in addition to HESA data. There is also clear policy appetite for this option.

If this option is not pursued, for some reason, then a link between ASHE and standalone HESA data should be considered instead. The benefit of this option would be maximised if HESA data from 1994-95 could be included in the linkage (rather than just data from the early 2000s, which is what ONS currently holds), but the benefits of acquiring this additional data would need to be considered alongside the time and resource costs involved in doing so.

Other UK nations

Wales is pursuing its own linkage between ASHE and administrative education data. Assuming their business case is approved, it would be sensible for the WED team (or another team taking forward a link between ASHE and LEO or HESA data in England) to keep in close contact with the team doing this work in Wales to ensure that any learning and/or benefits from the creation of similar derived variables are maximised.

Given the data developments occurring in Scotland and Northern Ireland, there would also be merit in considering a link between ASHE and HESA or LEO data in Scotland and Northern Ireland as well (as data availability permits), with data linkage teams across the UK working together to create comparable information on HE participants and graduates across the four nations as far as possible.

For comparable data to be of maximum benefit to researchers, however, one would ideally work towards making a single UK-wide ASHE dataset (linked to relevant education administrative (and other) datasets) available via one or more of the UK's secure research services. This is clearly a much more ambitious proposal that would need to form part of wider discussions around data sharing and linkage across the UK nations.

Suggested next steps

- Maintain a dialogue with officials at DfE and the devolved administrations regarding plans for linking ASHE and LEO and identify how support from ADR UK could be used to support and develop these plans.
- Keep in touch with representatives from other UK nations regarding current and future plans for linking ASHE to other data, to share learning and identify opportunities for future data development.
- Join or instigate discussions regarding the possibility of sharing data – including, but not limited to, linked ASHE data – across the four UK nations, to maximise the benefits of these linked data projects.

Acknowledgements

The author gratefully acknowledges funding from ADR England for this feasibility study.

The following individuals were kind enough to share their expertise and insights to inform the study:

- Thomas Alexis, Scottish Government
- Orla Bateson, Northern Ireland Statistics and Research Agency
- Chiara Cavaglia, Centre for Economic Performance, London School of Economics
- Alice Chaverot, Department for Education
- Rachael Colquitt, Office for National Statistics
- Dan Cook, Higher Education Statistics Agency
- Rhys Davies, Cardiff University
- Erin Early, UCL Centre for Education Policy and Equalising Opportunities
- Anthony Harris, Department for Education
- Zeeshaan Ishaq, Department for Education
- Jay Khamis, Department for Education
- Cecilia MacIntyre, Scottish Government
- Tej Nathwani, Higher Education Statistics Agency
- Dermot O'Reilly, Queen's University Belfast
- Jonathan Waller, Higher Education Statistics Agency
- Ben Waltmann, Institute for Fiscal Studies
- Nicholas Webster, Welsh Government
- Philip Whitham, Office for National Statistics

This report does not necessarily reflect the views of the individuals or organisations named, and the author bears sole responsibility for any inaccuracies or misinterpretations of the information reported.

References

Britton, J., Buscha, F., Dickson, M., van der Erve, L., Vignoles, A., Walker, I., Waltmann, B., Zhu, Y., (2020a), *The earnings returns to postgraduate degrees in the UK*, Department for Education Research Report 996.

Britton, J., Dearden, L., van der Erve, L., Waltmann, B., (2020b), *The impact of undergraduate degrees on lifetime earnings*, Institute for Fiscal Studies Report R167.

Britton, J., Dearden, L., Waltmann, B., (2021a), *The returns to undergraduate degrees by socio-economic group and ethnicity*, Department for Education Research Report 1100.

Britton, J., van der Erve, L., Belfield, C., Vignoles, A., Dickson, M., Zhu, Y., Walker, I., Dearden, L., Sibieta, L Buscha, F (2021b), *How much does degree choice matter?* Institute for Fiscal Studies Working Paper 21/24.

Naylor, R., Smith, J Telhaj, S., (2016), Graduate returns, degree class premia and higher education expansion in the UK, *Oxford Economic Papers*, Vol. 68, pp. 525-545.

Walker, I. Zhu, Y., (2011), Differences by degree: evidence of the net financial rates of return to undergraduate study for England and Wales, *Economics of Education Review*, Vol. 30, pp. 1177-1186.

Visit the [Wage & Employment Dynamics website](#)

Visit the [ADR UK website](#)

[@Wage_Dynamics](#)

[@ADR_UK](#)

