ADR ENGLAND

Data Explained

Longitudinal Education Outcomes (LEO)

Using linked National Pupil Database and Higher Education Statistics Agency data

Author: Dr Paul Martin
Date: January 2025

This Data Explained summarises experiences and learning from working with the National Pupil Database (NPD) and Higher Education Statistics Agency (HESA) linked datasets. These are two of the datasets which contribute to the Longitudinal Education Outcomes (LEO) data, although they can also be accessed separately, as they were for the project underpinning this article. This publication is intended to help guide future researchers using this data and to provide feedback into future dataset development and documentation.

The administrative data discussed in this Data Explained was made securely available to support a doctoral research project undertaken at the University of Warwick. The data used in this research project comes from the Department for Education and the Higher Education Statistics Agency and was accessed through the Office for National Statistics (ONS) Secure Research Service.

Introduction

The National Pupil Database (NPD) is a collection of datasets held by the Department for Education (DfE) in the UK Government. The NPD is one of the richest education datasets in the world. It contains the details of millions of young people who have passed through the English school and college system, with data available concerning pupils' attainment and personal characteristics. The first version of the NPD was produced in 2002. The DfE use the NPD to support accountability and school improvement, however, de-identified extracts of the data are also made available to researchers for the purpose of carrying out research in the public interest.

The Higher Education Statistics Agency (HESA) is part of Jisc, which is the Designated Data Body for higher education in England. HESA collects and processes data from higher education providers, including the HESA student record, which was accessed for this project. The HESA student record contains a range of variables including those relating to student registration status, level of study and institution attended.

This Data Explained document relates to use of linked NPD and HESA data which has been accessed following a direct application to the DfE[1]. As outlined above, NPD and HESA records also form part of the Longitudinal Education Outcomes (LEO) data, which additionally includes data on university applications, further education records, and tax and benefit records, including information about employers and the Covid-19 support schemes, and can be accessed via a different application process[2].

I used the linked NPD and HESA dataset to carry out a piece of research which investigated the extent to which the personal characteristics and area of residence of young people in England has a bearing on the likelihood of higher education participation. A summary of the outcomes of this research will be published in an upcoming Data Insight.

How is the data collected?

The NPD data used in this study was drawn from two sources – the pupil-level School Census and awarding bodies (sometimes referred to as 'exam boards').

The Pupil Level Annual School Census was first introduced in 2002 before being replaced by the School Census in 2006. State schools take part in the School Census by supplying the DfE with data concerning all enrolled pupils and their characteristics. Participation is mandatory and data is collected three times per year.
When school pupils in England undertake externally-accredited assessments – such as General Certificate of Secondary Education (GCSE) assessments, which are typically taken at age 16 – the

[1] Further details about how to apply to the DfE directly to access linked NPD and HESA data can be found at the following link, along with a copy of an application form:
https://www.gov.uk/guidance/apply-for-department-for-education-dfe-personal-data
[2] Further details about the LEO data, including how to access it, can be found here:
https://www.gov.uk/government/publications/longitudinal-education-outcomes-leo-dataset/longitudinal-education-outcomes-leo-data

awarding bodies who certify these assessments supply the DfE with pupil attainment data. Overall, the DfE receives data from approximately 150 different awarding bodies concerning which qualifications have been taken and how these have been graded. The DfE then matches this externally validated attainment data to their own records from the School Census.

All higher education providers have a statutory duty to submit data concerning their students to HESA. Providers collect data about their students during the application and registration process and HESA collects data from providers at the end of each academic year. HESA is itself obliged to share this student data with the DfE. The DfE is then able to link HESA student data with the NPD, using a process of 'fuzzy matching' where records are linked based on combinations of personal data such as names, postcodes and dates of birth.

When linked NPD and HESA data is supplied to researchers, separate NPD and HESA tables are provided. These tables contain pseudonymous pupil matching references which researchers can use to link records together for individual pupils/students. These matching references are generated by the DfE using the fuzzy matching method described above. This means that researchers are able to match individual records together themselves, without being supplied with personally identifiable data such as names, postcodes and dates of birth.

Key variables

In my research I used the Key Stage 4 attainment table from the NPD for the 2014-15 academic year. This contains details of pupils' attainment, as well as their characteristics, populated from the School Census. For the HESA data, I used the student records from both the 2017-18 and 2018-19 academic years. Below I describe the key variables used:

Data linkage: The NPD and HESA records were linked using the 'KS4_PupilMatchingRefAnonymous' variable from the NPD and the 'HE_PUPILMATCHINGREFANONYMOUS' variable from the HESA data tables.

Population selection: The population of interest in this research was all state school pupils in England who were 15 years old at the beginning of the 2014-15 academic year. The 'KS4_AGE_START' variable was used to identify and remove any pupils who were not aged 15 and the 'KS4_NEW_TYPE' variable was used to identify any remove any pupils who were not attending state schools.

Higher education data: The 'HE_XPSR01' variable was used to identify whether a student was registered at a higher education provider and the 'HE_XINSTID01' variable was used to observe which provider a student was registered at.[3]

[3] Note that in linked NPD-HESA data, institutions are identifiable, while in LEO data, institutions are pseudonymised. This is discussed in more detail later in the article.

Personal characteristics and attainment: The table below shows the variables that were used to observe pupils' personal characteristics and attainment:

NPD variable name	Description	Further details
KS4_FEMALE	Pupil gender	Shows whether or not the pupil is female (gender is a binary variable in the NPD)
KS4_ETHNIC	Pupil ethnicity	Shows pupil ethnicity in one of 98 possible categories
KS4_FSM	Eligibility for free school meals	A dichotomous variable showing whether a pupil is known to be eligible for free school meals during the year in which they took their KS4 exams (this is an indicator of low household income)
KS4_PPCODE	Pupil postcode	The pupil's home postcode – this was used to match in data concerning the characteristics of pupils' neighbourhoods
KS4_PTSCNEWE_PTQ_EE	Attainment points score measure	A points score based on the pupils' best 8 GCSE qualifications
KS4_EBACC_PTQ_EE	English Baccalaureate measure	An indication of whether a pupil has achieved a good pass in a range of traditional academic qualifications

It is worth noting that variable names can change over time in the NPD and HESA datasets. There is a helpful online search tool for the NPD which can help researchers to identify which variables may measure similar constructs in different years.[4]

What can the data be used for?

When used individually, both the NPD and HESA datasets can be used to answer a wide range of research questions, for example, about the variation in school and university outcomes for students from different backgrounds. Linking the datasets additionally opens up a whole new range of questions, such as which factors are associated with an increased (or reduced) likelihood of a young person participating in higher education.

For example, HESA data on its own can reveal what proportion of university students are of different ethnic backgrounds. However, this in itself does not show how likely young people of different ethnic backgrounds are to attend university, given that the relative proportions of pupils from different ethnic backgrounds within the wider population of young people are likely to be different.

[4] https://www.find-npd-data.education.gov.uk/

Similarly, HESA data does not contain information about a students' eligibility for free school meals during their school career, so it is not feasible to understand the disadvantage level of university students using this particular measure from HESA data alone.

Linking the entire cohort of NPD data with HESA thus enables questions such as the following to be answered:

- What percentage of all school pupils of different ethnic backgrounds (in a given age cohort) progress to university [by the age of 19], and how do these compare?
- What percentage of all school pupils eligible for free school meals (in a given age cohort) progress to university [by the age of 19], and how does this compare with the percentage of all school pupils not eligible for free school meals who progress?

Through the use of multivariate analysis, researchers can attempt to isolate the influence of particular characteristics once other average differences between groups of students are controlled for statistically. For example, the linked data can be used to investigate whether the higher HE participation rate of young women compared to young men can (or cannot) be solely attributed to the higher average school attainment of young women.

Given the wide range of variables available in the HESA dataset, the linked data can also be used to investigate the relationship between the attainment and characteristics of school pupils and the likelihood of other HE-related outcomes such as:

- Attending a more selective university
- Studying a particular subject at university
- Attaining a 'good' degree classification, such as a first or upper second class degree
- Leaving the family home to attend university
- Undertaking postgraduate study
- Withdrawing (or 'dropping out') from university study after enrolment

The availability of individual postcode data within the NPD enables new variables to be merged into the dataset (though postcode data is not available for independent school pupils). For example, in my research I used postcode data to introduce new variables which classified pupils' neighbourhoods according to both the Index of Multiple Deprivation[5] scale (which measures neighbourhood-level deprivation) and the POLAR[6] measure (which categorises areas according to rates of higher education participation). Researchers can also use the IDACI (Income Deprivation Affecting Children Index) neighbourhood-level measure which is already available in the NPD. Individual postcode data is not available in the version of the NPD made available in LEO, however in this case researchers may still be able to introduce new neighbourhood-level variables by merging these in using the lower super output area (LSOA) variable in the school census. This provides a more approximate measure of geographical location.

[5] https://www.gov.uk/government/collections/english-indices-of-deprivation
[6] https://www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/about-polar-and-adult-he/

Linked NPD-HESA data offers researchers certain opportunities which are not available to users of the LEO dataset, due to the fact that the two datasets are shared using different legal gateways. With linked NPD-HESA data, researchers are able to observe the identities of the particular universities, schools and colleges that students have attended. This may be useful when it comes to answering certain research questions. In contrast, in the version of the NPD and HESA datasets which are made available in LEO, the names (and reference numbers) of universities, schools and colleges are pseudonymised.

Similarly, with linked NPD-HESA data, it is possible for researchers to ask DfE to match in their own datasets at the individual level. For example, data concerning the participants of an initiative intended to encourage HE participation could be matched into the dataset, with a view to gauging the effectiveness of this initiative. Such matching must be carried out by DfE, since personally identifiable variables (such as names and dates of birth) are never shared with researchers. This type of matching is not currently possible with LEO.

Existing research or examples of previous research

Some existing research has analysed the relationship between the attainment and characteristics of school pupils and their likelihood of progressing to higher education using linked NPD-HESA data. Studies include:

- Chowdry, H., Crawford, C., Dearden, L., Goodman, A. and Vignoles, A. (2013), Widening participation in higher education: analysis using linked administrative data, *Journal of the Royal Statistical Society: Series A*, Vol. 176, pp. 431-457. https://doi.org/10.1111/j.1467-985X.2012.01043.x
- Crawford, C., & Greaves, E. (2015). *Socio-economic, ethnic and gender differences in HE participation*. Department for Business Innovation and Skills. https://www.gov.uk/government/publications/higher-education-participation-socio-economic-ethnic-and-gender-differences
- Gorard, S. (2018). Widening participation to higher education. In *Education policy, equity and effectiveness* (1 ed., pp. 147-170). Bristol University Press. https://doi.org/10.2307/j.ctv4rfrfg.17
- Britton, J. and van der Erve (2020), *Family background and access to postgraduate degrees*, Institute for Fiscal Studies Briefing Note. https://ifs.org.uk/sites/default/files/output_url_files/Family-background-and-access-to-postgraduate-degrees_.pdf

Data limitations encountered

In the case of this research project, the data application and approval process was lengthy, with more than 7 months elapsing between the submission of the data application and the availability of the data on the ONS Secure Research Service. Researchers should be mindful of this and allow plenty of time for data access as part of their project plans.

One limitation of the NPD data is the presence of duplicate entries in data tables, for example in the Key Stage 4 attainment table which was used in this research. The number of duplicate cases was small, for example out of 622,519 cases in the 2014-15 Key Stage 4 attainment table, only 519 cases did not have a unique pupil matching reference. However, this makes it challenging to match data across tables, which often relies on the linking variable(s) being unique. Duplicate cases can sometimes contradict each other but can also supplement each other, for example if a variable is missing in one duplicate case and present in another. Researchers will face a dilemma as to whether to simply discard duplicate cases (since there are so few of them) or to invest time in producing code which will merge duplicate cases whilst extracting the maximum amount of data from them.

The data available in the NPD is more limited for pupils attending independent (i.e. fee-paying) schools. This is because independent schools do not take part in the School Census, so only data supplied by awarding bodies relating to their attainment and a limited set of characteristics (e.g. gender) is available for these pupils. However, data availability overall is very good for pupils in state schools. For my research (which focussed on state school pupils), all of the data I required was fully available for 97.3% of all cases, meaning that bias associated with missing data is limited.

The NPD is limited to pupils attending schools in England, with no data available for pupils attending schools in the other nations of the UK. However, the HESA data does allow for the tracking of pupils to universities in Scotland, Wales or Northern Ireland as long as those pupils were originally domiciled in England.

Suggested improvements

- Unlike other education datasets, such as LEO and GRADE (Grading and Admissions Data for England[7]), the linked NPD-HESA dataset is not listed in either the ADR UK or ONS dataset catalogues, making the dataset harder to discover and cite.
- NPD data tables would be easier to use if there were no duplicate entries and only one row per pupil. DfE is presumably best placed to determine how to deal with these duplicates, and this would make it easier to replicate results across studies, as there will be fewer decisions for researchers to take regarding how to treat these observations.
- While not encountered specifically in my research project, research using multiple cohorts of data can face challenges in identifying the same variables across cohorts as a result of variable name changes. A look-up table enabling users to identify which variables capture the same information in different years would be very useful.

[7] GRADE links NPD to the detailed data from Ofqual underlying GCSE and A-level grades for cohorts taking exams between 2017 and 2020, and university applications data for 18 year olds from UCAS. For more information, see https://www.gov.uk/government/publications/grading-and-admissions-data-for-england-grade-framework/grading-and-admissions-data-for-england-grade-framework.

Suggested future data linkages

- Data in the NPD concerning pupils' socioeconomic background is limited. Any further data linkages which introduce variables concerning more finely-grained measures of household income, measures of parental occupation or parental level of education would be welcome. Alternatively, HESA data might be linked to other datasets which feature richer household-level data, such as the Growing up in England dataset.
- As the linked NPD-HESA data is often used for research relating to widening participation in higher education, linkages with databases of widening participation outreach activity (such as the HEAT dataset) would also be welcome.

Recommendations to data owners

- Documentation relating to the dataset – such as a PDF user guide – may aid prospective users of the dataset in judging whether they will be able to use the data to answer their research questions. This could complement the existing "find and explore NPD" online tool[8].
- Given that it can take a long time to complete the data application form, researchers may benefit from a short grace period being applied in cases where an application form is submitted only shortly after the application form template on the website is updated.
- Similarly, it can be challenging for researchers to determine precisely the variables they need to address their research questions before they start work on the project, including which external datasets they would benefit from linking in. Making it easier for researchers to amend their applications without having to go through the entire approval process again from scratch would enable more research in the public benefit to be completed more quickly and to be of higher quality.

[8] https://www.find-npd-data.education.gov.uk/

Disclaimer

Parts of this work were produced using statistical data from the Office for National Statistics (ONS). The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce ONS aggregates.

Acknowledgements

This research was undertaken at the Department for Education Studies at the University of Warwick.

Contact

Name: Dr Paul Martin
Email: paul.e.martin@ucl.ac.uk

ADR ENGLAND

UKRI Economic and Social Research Council