Linked Education Administrative Datasets (National Pupil Database), England

USER GUIDE

Version 1.0, November 2020

Understanding Society

Economic and Social Research Council

Contents

1. Introduction

Data linkage is one of the key innovative features of *Understanding Society*, the UK Household Longitudinal Study (UKHLS), which allows researchers to develop and implement new research agendas. It includes linkage to administrative datasets in key policy areas as well as other external data sources. *Understanding Society* currently pursues data linkage in a wide range of topic areas, including education, health, economic circumstances, and transport as well as area characteristics.

Education administrative records are collected separately in each of the four countries of the UK. This user guide covers education records collected in **England** as part of the National Pupil Database (NPD). Education linkage with Scotland and Wales are underway and will be released separately. The NPD combines the examination results of pupils with information on pupil and school characteristics and is an amalgamation of a number of different datasets, including Key Stage attainment data and Schools Census data. This user guide gives an overview of how the NPD has been linked to *Understanding Society* and what data are available.

2. Education Data Linkage England

2.1 Consent to education data linkage

Consent to education data linkage was collected separately for two groups of survey members, adults who are young enough to have educational records included in the National Pupil Database[1] and children aged 4-15. While young adults (16 years and older) were able to consent on their own behalf, responsible adults consented to data linkage on behalf of their children aged under 16.

Consent to linkage of education data is collected in *Understanding Society* at intervals of generally 3 years. The first consent was collected at Wave 1 of UKHLS, the second at Wave 4.[2] At both waves the consent procedure consisted in handing out an information leaflet to respondents, asking them a consent question as part of the main survey and asking them to sign a consent form (for children this was signed by a responsible adult on behalf of the child). At Wave 4 the consent question was worded differently depending on whether respondents had already consented at Wave 1 and were therefore re-affirming a previous consent, or whether they had refused at Wave 1 or had not previously been asked for consent (for example because their child/children had not reached school-age at Wave 1, they were new to the study, had turned 16 or had missed the consent question). The consent questions can be found in the Wave 1 and 4 questionnaires https://www.understandingsociety.ac.uk/documentation/mainstage/questionnaires and the information leaflet and consent forms can be found in the Wave 1 and 4 fieldwork documents https://www.understandingsociety.ac.uk/documentation/mainstage/fieldwork-documents. The

[1] The earliest birthdate that can be matched in the National Pupil Database is 1. September 1983.
[2] Consents were also collected at Wave 7 and these will form the basis of future linkage.

Wave 4 consent module was also carried in Waves 5 and 6 for new entrants into the study and for 'rising 16s' – children who turn 16 and are eligible to give their own consent.

Respondents can withdraw their consent to link their education records at any time. If at the time of withdrawal the data has not yet been linked, the individual is removed from the set of consenting individuals. If linkage has already taken place, no further data will be added for that individual.

Table 1 gives an overview of the population eligible to receive the consent question at Waves 1 and 4, separately by whether the consent was for a child or an adult, and by gender, and the number and proportion of respondents who consented to education data linkage. Consent rates were higher for adults than for children at both Waves 1 and 4. The population of adults for whom data is available in the NPD increased between Waves as more children move into that age range.

Table 1: Consent and match rates

	Children aged 4-16			Adults 16+		
	All	girls	boys	all	women	men
Wave 1						
Eligible population	9,745	4,743	5,002	5,041	2,758	2,283
Consented	6,480	3,131	3,349	3,915	2,141	1,774
% of eligible	66.5	66.0	67.0	77.7	77.6	77.7
Matched	5,331	2,608	2,723	2,193	1,174	1,019
% of consent	82.3	83.3	81.3	56.0	54.8	57.4
% of eligible	54.7	55.0	54.4	43.5	42.6	44.6
Wave 4						
Eligible population	7,617	3,697	3,920	6,664	3,599	3,065
Consented	4,739	2,344	2,395	5,260	2,892	2,368
% of eligible	62.2	63.4	61.1	78.9	80.4	77.3
Matched	4,678	2,311	2,367	4,700	2,545	2,155
% of consent	98.7	98.6	98.8	89.4	88.0	91.0
% of eligible	61.4	62.5	60.4	70.5	70.7	70.3

2.2. Linkage process

For the data linkage the Institute for Social and Economic Research (ISER) generated an anonymised ID for all consenting individuals and extracted Forename, Surname, Date of Birth, Postcodes and Address details from our respondent database. At Wave 4 the consent question covered permission to use school identifiers for matching to improve matching rates particularly for young adults who no longer live at the same address as that contained in school records. Therefore, from Wave 4 school codes were used as a further matching variable. Anonymised ID and matching variables were transferred securely to the Department for Education (DfE) where the data were matched using different combinations of the identifying characteristics.

At Wave 1, stage 1 matching included exact matches to Pupil Level Annual School Census (PLASC) data sets using names, date of birth, gender and postcode (63% of matches); stage 2 used fuzzy name matches to Census, plus date of birth, postcode, gender check (4.8% of matches); stage 3 used fuzzy postcode matches to Census plus names, date of birth, postcode, gender check (4.7% of matches). Individuals who were matched at Wave 1 and re-consented at Wave 4 were re-matched at Wave 4.

At Wave 4 forename, surname, date of birth, postcodes, address details and school code (LAESTAB) were used for matching. All current and historical data in NPD were searched. Stage 1 matching included exact matches on names, date of birth and postcode (75% of matches); stage 2 used names, date of birth and partial postcode (9.0% of matches); stage 3 used fuzzy names, date of birth and postcode (8.6% of matches); stages 4-8 used further combinations of matching variables (6.2% of matches).

The achieved match rates for adults and children, by wave matched and gender are shown in Table 1. Match rates were low among young adults at Wave 1 but improved at Wave 4, possibly because school attended was used as an additional matching variable.

3. Available Data

3.1 Overview of NPD data

The NPD is a register database of all pupils in state schools in England. It contains attainment data as children progress through school, as well as other rich information, including on pupil background, absences and exclusions from school. The Table gives an overview of the types of information covered in the data files and the ages it relates to.

Table 2: NPD data files

Name of data file	Description	Age	Content	First year collected
EYFSP	Early Years Foundation Stage Profile	5	Attainment data based on teacher assessments collected in primary schools at the end of Reception year. New measures introduced in 2012/13.	2002/03
KS1	Key Stage 1	7	Attainment data based on teacher marked tests collected in primary schools at the end of year 2. New measures introduced in 2015/16	1997/98
KS2	Key Stage 2	11	Attainment data from national, externally marked tests taken at the end of primary school (end of year 6). National Curriculum Levels were replaced by scaled SATS scores in 2015/16.	1995/96
KS3	Key Stage 3	14	Attainment data from national curriculum assessments and teacher assessments at the end of year 9 which ended in 2007/08 and 2012/13 respectively.	1997/98

KS4	Key Stage 4	16	Attainment data based on national exams in General Certificate of Secondary Education (GCSE) and equivalent qualifications at the end of secondary school (year 11). Changes in grading introduced from 2016/17. Data contain grades in many subjects as well as summary indicators of attainment.	2001/02
KS5	Key Stage 5	17,18	Attainment data based on post-16 assessment in school sixth forms and FE colleges. Includes A-levels, AS-levels and equivalent results.	2001/02
Census	Pupil Level Annual School Cenus	4-18	School Census data collected in January from pupils in state schools. Includes background characteristics.	2001/02
Absences		4-18	Information on pupil absences by term and annually by reason for absence.	2005/06
Exclusions		4-18	Information on pupil temporary and permanent exclusions from school, including reason for each exclusion.	2005/06

3.2 Sample and data in the current release

The bulk of the data contained in the current release was linked in 2018 based on consents collected in Wave 4 and contains all NPD records up to academic year 2017/18. In addition, NPD data for individuals who had consented and were successfully linked at Wave 1 but not matched again at Wave 4 (due to non-response, exiting the survey, withdrawing consent) was added to the current release. This was linked in 2012 and contains NPD records up to academic year 2012/13. We have permission to continue to use this data but not to update it to include further academic years. Each of the NPD files contains a variable 'link_indicator' which indicates whether a data point originates from wave 1 or wave 4 linkage.[3]

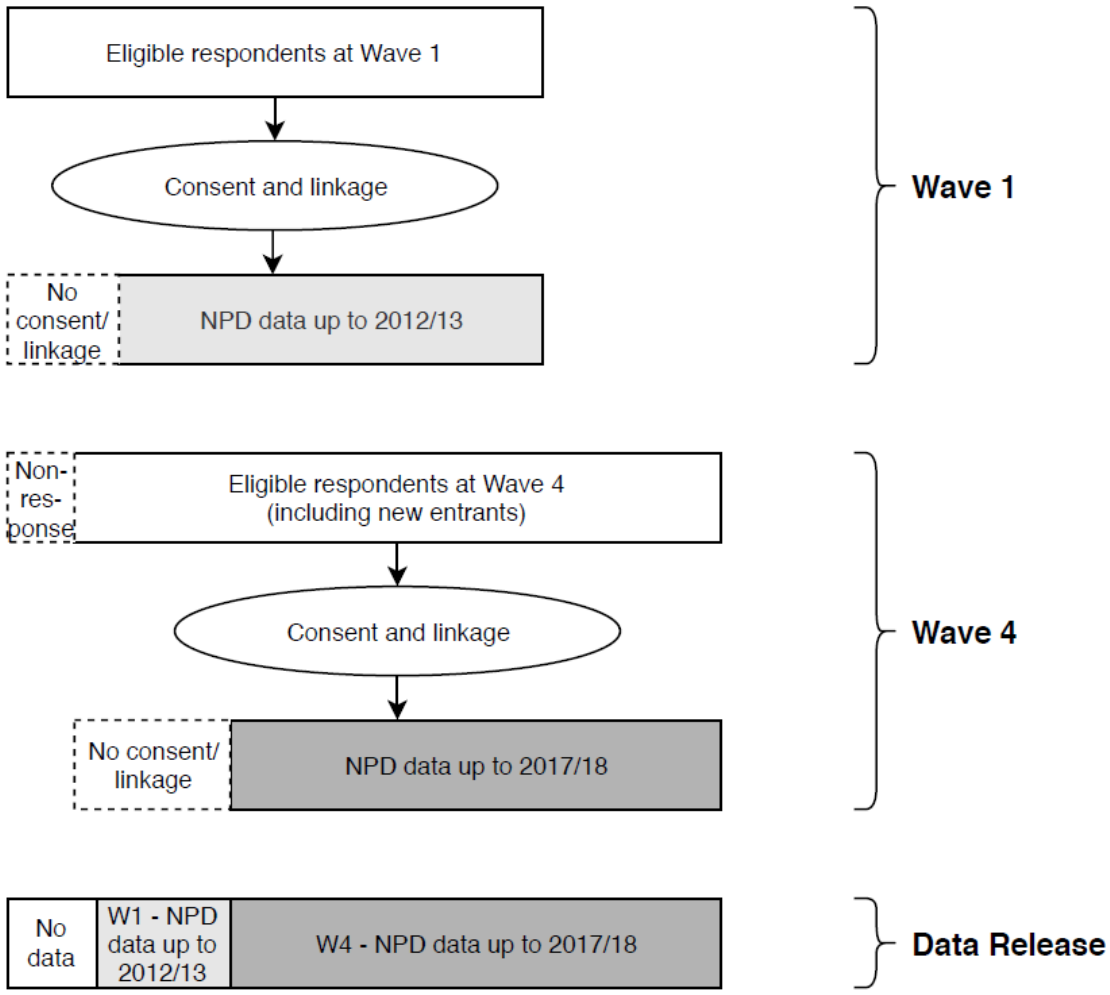

Figure 1: Sample included in the data release

[3] Note that in the 2012 linkage exercise the most recent Pupil Level Annual School Census record was returned while in the 2018 linkage each annual record is included in the data.

Figure 1 gives an overview of the eligible respondents and linked sample for Waves 1 and 4, and of the sample included in the current release. Data from Wave 1 linkage is only included in the current release for individuals who were linked at Wave 1 and could not be linked again at Wave 4. The match performed at Wave 4 overwrites any match performed at Wave 1. We have no linked data in the current release for individuals who never consented.

Data

Once a link was established, i.e. a UKHLS respondent or child was identified in the NPD, all the available records from the NPD files described in Table 2 above were extracted for that individual. How far back in time records can be added will depend on the person's birth year and the year each record type was first collected. Moreover, for adults all the records up to age 18 can be added while children who are still in school are accumulating more data over time.

The NPD data files are intended to be linked to the main UKHLS survey using the pidp identifier included in each file. Three options are available depending upon the level of data available (in ascending order of access restrictions): SN 6614 (End User Licence), SN 6931 (Special Licence) or SN 6676 (Secure Access). Please refer to the documentation for each study for further details. Each file of this dataset also includes an academic year indicator to indicate the year the data refers to.

As a result of the data structure the number of observations will vary between NPD data files. All the frequencies and summary statistics, as well as descriptors of the variables can be found online: https://www.understandingsociety.ac.uk/documentation/linked-data/education-data-linkage/npd-variables

To give an overview of the available sample sizes Table 2 lists the number of observations in each of the NPD data files, as well as the number of unique individuals, observations per school, local authority and per academic year.

Table 3: Descriptives for matched NPD data

Data set	Observations (N)	Unique individuals (I)	N per school (I per school)	N per Local Authority	N per academic year
EYFSP	4,843	4,843	1.6 (1.6)	32.5	302.7
KS1	9,933	9,933	2.0 (2.0)	64.9	473.0
KS2	10,824	10,824	2.0 (2.0)	69.4	470.6
KS3	7,799	7,709	2.9 (2.9)	51.0	487.4
KS4	7,245	7,206	2.6 (2.5)	47.4	426.1
KS5	9,166	4,853	5.3 (2.8)	58.8	539.2
Census	70,136	13,421	7.7 (1.5)	461.4	4,125
Absences	63,559	11,578	7.8 (1.4)	415.4	3,739
Exclusions	1,503	916	n/a	n/a	125.3

3.3 Weights for linked data

The consent and linkage process potentially leads to non-random selection of individuals for whom linked data is available. We are working on guidance on deriving analysis-specific weights which will cover issues specific to linked data. Guidance will be made available soon.

4. Data access

The linked data is available from the UK Data Service, the details for which can be found at: https://discover.ukdataservice.ac.uk/catalogue/?sn=7642. Due to the sensitive nature of the data it is classified as Secure Access or Controlled data. This classification means that the data can only be accessed through the UK Data Service Secure Lab. Full details of the access requirements and the application process can be found at: https://www.ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab.aspx. It should be noted that access is restricted to researchers registered at a UK institution. In addition, before the data can be accessed, researchers must have attended a Safe Researcher training course. One consequence of this is that the time from applying for the linked data to having access to it can be much longer than for other types of data. If time is an issue, or if you have other queries regarding access to the data then please get in contact with the UK Data Service in the first instance: https://www.ukdataservice.ac.uk/help/get-in-touch.aspx.

5. Citation

If you use *Understanding Society* data you must cite every study that you use.

The bibliographic reference for this study is as follows:

Department for Education, University of Essex, Institute for Social and Economic Research. (2020). Understanding Society: Linked Education Administrative Datasets (National Pupil Database), England, 1995-2018: Secure Access. [data collection]. 3rd Edition. UK Data Service. SN: 7642, http://doi.org/10.5255/UKDA-SN-7642-3.

All works which use or refer to these materials should acknowledge these sources by means of bibliographic citation. To ensure that such source attributions are captured for bibliographic indexes, citations must appear in footnotes or in the reference section of publications.

Citing this User Guide

When citing this User Guide you can use the citation of this particular version quoted below.

Department for Education, Institute for Social and Economic Research (2020), *Understanding Society: Linked Education Datasets (National Pupil Database), England, 1995-2018, User Guide, Version 1.0, November 2020*, Colchester: University of Essex.