Office for National Statistics

Growing Up in England (GUiE)

Linking demographic, geographic, and household information to the highest level of educational attainment and vulnerability characteristics using the Growing Up in England (GUiE) dataset.

Contact:
Clare Melson
adrcuration@ons.gov.uk
+44 30 0067 2289

Release date:
18 October 2022

Next release:
To be announced

Table of contents

1 . Main points

- The Growing Up in England (GUiE) dataset provides users with detailed information about the educational journey of a cohort of children.
- The data offer a new level of insight into children who are vulnerable because of their circumstances, such as children from jobless households.
- GUiE can be used to research the barriers and gateways to improved educational outcomes and social mobility.
- This methodology describes the datasets included within the GUiE longitudinal dataset, and details the methodology used to link the 2011 Census to the feasibility All Education Dataset for England (AEDE); see Section 4 and Section 5 for more information.
- The richness and population size of the linked datasets enable researchers to perform sub-group analyses, like the sort conducted in our analysis; see Section 6 for more information.
- Our analysis uses personal and household characteristics, educational attainment, and vulnerability information to produce data tables, which are available in our accompanying datasets.

These are not official statistics and should not be used for policy or decision making. They provide users with information on how to use the Growing Up in England (GUiE) dataset to understand the educational journeys of children and young people.

2 . Background

Growing Up in England (GUiE) is a linked administrative dataset which was developed by the Office for National Statistics (ONS) in partnership with Administrative Data Research UK (ADR UK).

The ONS is the largest producer of official statistics in the UK, responsible for collecting, analysing, and disseminating statistics about the UK's economy, society and population. This information has traditionally come from surveys or the census; however, these sources often cannot provide enough detail or lack timeliness. For this reason, we are developing our use of administrative and linked data. See more information about the launch of the Administrative Data Research Partnership.

ADR UK is a partnership changing the way researchers access the UK's public sector data to enable better informed policy decisions that improve people's lives. ADR UK is formed of four national partnerships (ADR England, ADR Scotland, ADR Wales, and ADR Northern Ireland), and the Office for National Statistics (ONS). ADR UK aims to create a sustainable body of knowledge about how our society functions by linking data from across government.

The ADR UK partnership also facilitates safe and secure access for accredited researchers to these new linked datasets. ADR UK-funded research projects are tailored to decision makers' needs, to provide answers required to solve policy questions. This background and objective is summarised in ADR UK's mission statement: "Administrative data is an invaluable resource for public good. We're using it." See further information about Administrative Data Research UK (ADR UK).

GUiE aligns with ADR UK's overall goal of providing research-ready linked administrative data that support research for the public good.

The attainment gap between socio-economically advantaged and disadvantaged children indicates how socioeconomic status can impact upon a child's opportunities for future success in life. There is a need for research that can:

- enable us to understand the barriers and gateways to improved educational outcomes and social mobility
- inform the public policy that affects disadvantaged children and young people
- help to bridge the widening attainment gap

GUiE links administrative education data to census data, enabling research into the interaction between personal, household, and geographical factors and educational attainment. The richness of this linked data facilitates analysis of small subgroups of the population, who were previously underrepresented because of data gaps and sample size limitations.

The feasibility to create the GUiE linked dataset was previously assessed in a Proof of Concept (PoC) study, published in July 2020. It was concluded that administrative education data - sourced from the feasibility All Education Dataset for England (AEDE) - can be linked to the 2011 Census with a high linkage rate. Further details can be found in our Educational attainment and household composition, feasibility research and methodology.

Following this PoC, the feasibility AEDE was re-linked to census data in a new linkage environment. GUiE Wave-1 links education data from the feasibility AEDE for the academic years 2001 to 2002 to 2014 to 2015 to the 2011 Census. This enables information about students' highest level of educational attainment to be linked to their demographic, geographic, and household information from the census. Wave-2 joins additional information on vulnerability measures to this dataset, including information on absences, exclusions, Free School Meals (FSM), Children in Need (CIN), and Children Looked After (CLA).

3 . What is included in Wave-1 and Wave-2 of Growing Up in England (GUiE)

Growing Up in England (GUiE) links demographic, geographic and household information to data on highest educational attainment and vulnerability characteristics across a longitudinal cohort. This enables research into how a child's characteristics, and the characteristics of their household, could influence attainment. Educational attainment data were obtained from the feasibility All Education Dataset for England (AEDE). The addition of vulnerability data in Wave-2 enables unique insight into how educational attainment, sociodemographic data, and geography are linked to childhood vulnerability. This includes, for example, data relating to physical disabilities or children in care.

Demographic, geographic, and household information was sourced from the 2011 Census for England and Wales.

Individuals aged between 10 and 25 years on 31 August 2011 are included within the GUiE data if they were enrolled in government funded education, or non-government funded further education, in England between the academic years 2001 to 2002 to 2014 to 2015. However, some attainment data from independent schools are also present within the National Pupil Database (NPD), where students have taken regulated qualifications.

To create the GUiE data, linkage was done in two steps resulting in two "waves" of data.

Wave-1 was created by linking records from the feasibility AEDE to records from the 2011 Census in England and Wales. This first wave connects personal, familial, and household characteristics with educational attainment information. Wave-1 contains education data from the academic years 2001 to 2002 to 2014 to 2015.

Wave-2 was created by joining five vulnerability datasets to the Wave-1 dataset. These five vulnerability datasets were:

- absences
- exclusions
- English School Census (ESC)
- Children in Need (CIN)
- Children Look After (CLA)

Wave-2 datasets are not contained within the feasibility AEDE. They are joined to the Wave-1 data using the unique pupil matching reference, which is contained within the feasibility AEDE and Wave-2 datasets. Although absences and exclusions data are collected within the ESC, we refer to absences and exclusions as separate datasets.

Wave-2 vulnerability data are available for the academic years 2010 to 2011 to 2014 to 2015. Therefore, the fullest coverage of the GUiE dataset spans from 2010 to 2011 to 2014 to 2015.

2011 Census for England and Wales

The census takes place every 10 years and provides us with a picture of all the people and households in England and Wales. It provides a snapshot of family composition as well as demographic and socioeconomic characteristics for almost the entire population. Though the census provides only a snapshot, many personal characteristics, such as country of birth, are valid over time and remain relevant beyond the date of the census.

The 2011 Census contains a range of valuable information and includes data on:

- household composition
- ethnicity
- main language spoken
- country of birth
- religion
- qualifications
- occupation

See more information about the 2011 Census, including published information on how the Office for National Statistics (ONS) processed the information.

Though the feasibility AEDE was linked to the full 2011 Census for England and Wales, GUiE only contains records for individuals who were matched to the feasibility AEDE. Therefore, census information for other household members, for example parents and grandparents, is not available within GUiE unless these other household members are contained within the feasibility AEDE.

The census data do contain some household-level derived variables and demographic and socioeconomic information on the household reference person. For example, National Statistics Socio-economic classification (NS-SEC) of household reference person and number of unpaid carers in household are available within the census data provided within GUiE.

The reference person is the member of the household who is listed as the owner or renter of the accommodation, or who is responsible for the accommodation. If there are joint members of the household who are listed as the accommodation owner or renter, the individual earning the highest income is taken as the household reference person.

More information about the variables available within the GUiE data are available on the Secure Research Service (SRS) Metadata Catalogue, under the "data dictionary" tab.

Record swapping was carried out on the 2011 Census as a form of disclosure control. Record swapping is a method where "noise" is introduced for a small number of records. Specifically, it involves identifying records that are at most risk of statistical disclosure and matching them to other households with similar characteristics in nearby geographic areas. Information from these record pairs is then swapped. Record swapping is carried out so that the total number of records remains the same, but the records are protected from statistical disclosure. The number or percentage of records, the types of protected characteristics, and the nature of the variables that were subject to record swapping cannot be provided; providing this information increases the likelihood of statistical disclosure.

Since the publication of the Proof of Concept (PoC) in 2020, another census has been undertaken. Discussions are underway to determine how future GUiE iterations can include data from Census 2021.

Feasibility All Education Dataset for England (AEDE)

Created by the Department for Education (DfE), the feasibility AEDE is a large, longitudinal, record-level education dataset. It was initially supplied to the ONS to enable us to investigate the potential for administrative data to provide information on educational qualifications.

The feasibility AEDE was created from the National Pupil Database (NPD), Individualised Learner Records (ILR) data, which include Further Education (FE) data, and Higher Education Statistics Agency (HESA) data. It covers predominantly government-funded education in England from 2001 to 2002 to 2014 to 2015. Note that data for the full range of years are not available for all datasets. For example, ILR data are not available for the academic year 2001 to 2002.

The viability to create the feasibility AEDE was previously demonstrated, and details can be found in the published Proof of Concept (PoC) report. See further information about the feasibility All Education Dataset for England (AEDE), but that within the GUiE data, the feasibility AEDE does not contain HESA data.

All personal identifiers in the feasibility AEDE held by the ONS are pseudonymised (made non-identifiable) to ensure confidentiality. The method used ensures sensitive personal identifiable information (PII) is not revealed but can be used for data linkage.

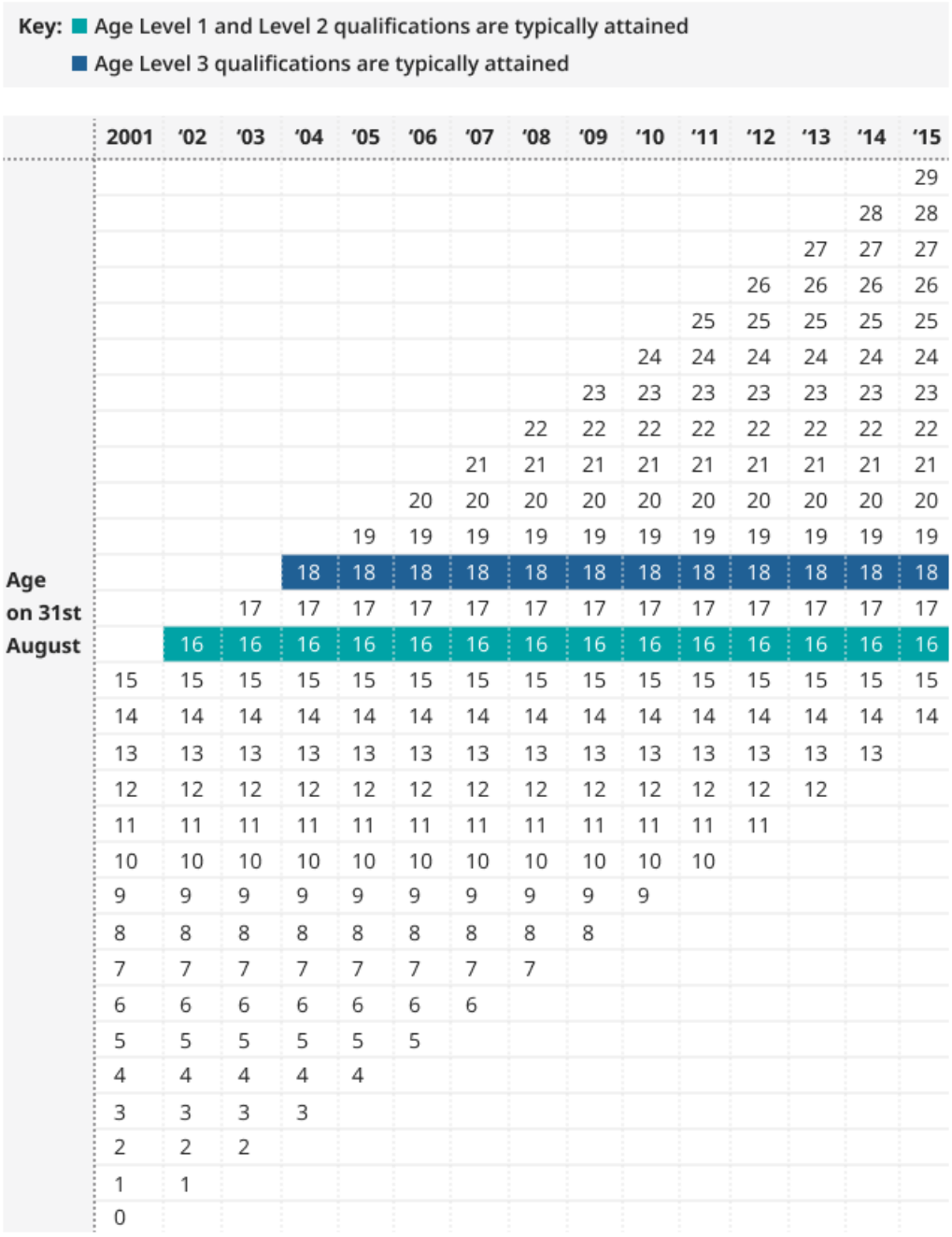
Students are included in the feasibility AEDE if they were enrolled in government-funded education or non-government funded FE in England in any academic year between 2001 to 2002 and 2014 to 2015, and aged between 10 and 25 years on 31 August 2011.

Note that this coverage may particularly affect level 3 data as students may transition from a government-funded education provider upon completing level 2, to a provider not captured within the feasibility AEDE.

A representation of the age groups included in the feasibility AEDE is shown in Figure 1.

Figure 1: Age range of the cohort contained in the feasibility AEDE

Age range of the cohort contained within the feasibility All Education Dataset for England and the age this cohort would typically attain level 1, 2, and 3 qualifications, England, 2001 to 2002 to 2014 to 2015



Source: Office for National Statistics - Growing Up in England (GUiE) data

GUiE links 2011 Census data to the feasibility AEDE, so the linked GUiE dataset covers the same academic years as the feasibility AEDE. Students are contained within the GUiE dataset if they have a record within the feasibility AEDE which has been linked to their 2011 Census record.

Further details on linkage methodology, results and quality can be found in Section 4: Wave-1 linkage methodology and Wave-2 join.

National Pupil Database (NPD)

The NPD is an administrative datastore that is held by the DfE. Within the feasibility AEDE, NPD data spans the years 2001 to 2002 to 2014 to 2015 and includes data from the Young Person's Matched Administrative Dataset (YPMAD) and the English School Census (ESC).

The full NPD held by DfE contains information related to all key stages. However, the extract of the NPD used within the feasibility AEDE is cumulative and only contains information on highest educational qualification. Moreover, young people who have never interacted with government-funded education are not included in the NPD, such as those who have only attended independent schools or have been electively home-educated.

The YPMAD is a derived dataset which enables the tracking of learner attainment in England. The YPMAD is generated by matching existing data sources together at an individual level using personal identifiers. It is available within the NPD but is not widely used or requested outside of the DfE.

The YPMAD contains information on students' highest educational qualification up to age 20 years. The data draw on attainment at levels 1, 2 and 3. Level 1, 2 and 3 qualifications are typically achieved during Key Stage 4 (KS4) and Key Stage 5 (KS5), but can also be achieved outside of these key stages.

The YPMAD data are cumulative, meaning that attainment does not have an associated key stage indicator. Therefore, while it is possible to use age as a proxy for key stage, it is not possible to use a key stage indicator to determine whether a qualification is obtained in KS4 or KS5. For example, it is not possible to determine whether a child attained a "Level 3" qualification in KS4 or KS5 using a key stage indicator.

The YPMAD is split into snapshot indicators and chronological indicators, which track those between the academic ages 15 to 20 years. For snapshot indicators, a learner appears once relating to that point in time. Snapshot indicators include:

- gender
- ethnicity
- binary flags which show study and achievement at each age
- geographic information about a learner's main provider

For chronological indicators, each learner has a record for each academic year. Measures included in chronological indicators are:

- highest qualification and provider where it was studied
- best attainment over time up to a given year
- participation in a given academic year (highest level studied, mode of study)

Many snapshot indicators are rolled forward into chronological indicators each year.

Not all of these snapshot and chronological indicators are included within the GUiE dataset. For more information about the variables included within the GUiE dataset, please see the SRS Metadata Catalogue.

Students' socio-demographic characteristics are contained within NPD attribute data. These characteristics are obtained from a blend of data sources including the spring term ESC, pupil referral units, and alternative provision censuses. These socio-demographic characteristics are then linked to attainment data recorded by awarding bodies.

Only the spring term ESC data are used to obtain students' socio-demographic characteristics within the NPD data. Autumn and summer term ESC data are available within Wave-2.

Individualised Learner Record (ILR)

The ILR database records socio-demographic characteristics and information on the learning aims of individuals in further education and work-based learning in England and attainment information. It spans the years 2002 to 2003 to 2014 to 2015.

ILR data are primarily used to underpin funding and commissioning decisions. Providers of further education in England must return ILR data for learners that receive government funding, even if they have only attended one episode of learning. The ILR contains information on learners from publicly-funded colleges, training organisations, local authorities, and employers acting as Further Education (FE) providers must collect and return every year.

FE colleges must send education data for all learners, including those not funded by government. Therefore, some non-government-funded learners are included in ILR data, for example, those undertaking learning subcontracted-in to college by local authority on behalf of another training provider.

All higher education institutes must return ILR data for learners through Advanced Learner Loans; adults who receive loans to cover tuition fees for a range of courses and access to diplomas of higher education.

The ILR data also contain Learning Aims Data (LAD), Learning Aims Reference Application (LARA) data and Learning Aims Reference Service (LARS) data. These datasets provide additional reference data about learning aims.

Wave-2 vulnerability data

Information on vulnerability characteristics from the ESC has been included within GUiE Wave-2 data. Vulnerability data are available for the academic years 2010 to 2011 to 2014 to 2015.

English School Census (ESC)

The ESC is a collection of pupil- and school-level information. The data are collated by local authorities into electronic returns and submitted to the DfE via a secure online data transfer system. There is published information on School Census: Data quality and processing (PDF,147KB).

Data are collected on a termly basis, resulting in three school censuses per academic year: autumn, spring, and summer. Note that the socio-demographic characteristics within the NPD are obtained from the spring term ESC, but in GUiE Wave-2 ESC data are available for spring 2010 to 2011 to 2013 to 2014, summer 2010 to 2011 to 2013 to 2014, and autumn 2014 to 2015.

The ESC contains data from:

- secondary schools
- middle-deemed secondary schools
- local authority-maintained special schools
- non-maintained special schools
- academies including free schools
- studio schools
- university technical colleges
- city technology colleges

Primary schools participate in the ESC, but they are not in the scope of the feasibility AEDE. Service children's education schools (schools for children of HM Armed Forces and Ministry of Defence (MOD) personnel) may participate on a voluntary basis.

The ESC does not include data from independent schools, and only collects information about individuals attending state-funded schools in England. The information on these individuals includes:

- ethnicity
- mobility
- free school meal eligibility
- special educational needs
- absences
- exclusions

Absences' data are available with termly coverage as part of the ESC for the academic years 2010 to 2011 to 2014 to 2015. These data contain information on authorised and unauthorised school absences and persistent absenteeism, including types and frequencies of absences.

Exclusions' data are available with termly coverage as part of the ESC for the academic years 2010 to 2011 to 2014 to 2015. These data cover permanent exclusions and suspensions (formerly fixed-term exclusions) and includes information on:

- year of exclusion
- reason for exclusion
- whether the child had been excluded on more than one occasion

Free School Meal (FSM) eligibility is recorded at multiple timepoints over the period of a child's education. The ESC collects information on whether a child was eligible for FSM on census day. It also contains information on whether a child has been eligible for FSM in the last three years, six years, or at any time over the course of their education.

The ESC also collects information on disability and Special Educational Needs (SEN), including type of disability and type of SEN support.

Children in Need (CIN) and Children Looked After (CLA)

CIN data are collected by local authorities and covers the years 2010 to 2011 to 2014 to 2015. A "child in need" is defined as "a child who is unlikely to reach or maintain a satisfactory level of health or development, or their health or development will be significantly impaired without the provision or services, or the child is disabled". This dataset provides information on:

- referrals to children's social care services
- assessments carried out on children referred to children's social care services
- children who are the subject of child protection plans

CLA data cover the years 2010 to 2011 to 2014 to 2015 and contain information on every child who is looked after by a local authority in one specific year. They also contain information on children who have been looked after for at least 13 weeks after they reach age 14 years.

4 . Wave-1 linkage methodology and Wave-2 join

Dimensions of quality

To ensure a broad understanding of our work and quality we adhere to the Code of Practice for Official Statistics (PDF, 420KB).

Overview

For Growing Up in England (GUiE) Wave-1, the feasibility All Education Dataset for England (AEDE) was linked to 2011 Census using deterministic and probabilistic linkage. Clerical review was then used to determine a cut-off threshold. We created 91 matchkeys, containing different combinations of personal identifiable information, and linked census records to the feasibility AEDE. The remaining unlinked records were probabilistically linked.

For Wave-2, Wave-2 data were joined to Wave-1 data using the person identifier (ID) present in the feasibility AEDE and the vulnerability data.

Feasibility AEDE processing

Before describing how the feasibility AEDE data were linked to 2011 data, we will detail the feasibility AEDE processing. More information about how the feasibility AEDE was processed is contained within our previous Proof of Concept (PoC) publication.

First, the feasibility AEDE attributes containing the source and academic year variables were linked to the feasibility AEDE index on the Pupil Matching Reference (PMR) number. The PMR gives each pupil a pseudonymised identifier, which is unique to them and allows matching across datasets without giving away their identity. The purpose of this step was to get the Unique ID from the feasibility AEDE index onto the attributes, which is needed to link to the matchkey file.

Once the feasibility AEDE attributes had been linked to the feasibility AEDE index, the index was subset into the 2010 to 2011 academic year where ESC was identified as the source dataset. Multiple entries of individual records have been removed; this process is known as de-duplication which was completed on the PMR and using a nodupkey procedure in Statistical Analysis System (SAS). The nodupkey procedure retains only the first instance of a record.

Duplicated records were removed. These individuals were then linked to the original matchkey file on the PMR to extract the correct matchkey records for linkage.

The resulting feasibility AEDE matchkey file, created for linkage, contained over 161 million records covering the academic years 2000 to 2001 to 2014 to 2015. For the purposes of this feasibility AEDE-census linkage, the file was subset into the ESC records for the 2010 to 2011 academic year, because this was the year closest to the 2011 Census. This is important because the information collected in this year is most likely to match that of the census, particularly for addresses. Overall, this increases the number of records which will link.

The feasibility AEDE was linked to the 2011 Census within our PoC, however this linkage was performed using hashed data in our legacy processing environment. GUiE Wave-1 linkage was conducted using data in the clear in the ONS's processing environment, the Data Access Platform (DAP).

Linkage in the clear

For the PoC, data were linked using hashed data, where raw identifiable data is transformed into a unique string of letters and numbers. The nature of the hashing process means that in cases where there are spelling errors or inconsistencies between two records relating to an individual, the hash values will not be identifiable as being similar. The linkage for GUiE Wave-1 was conducted in the clear, using identifiable data. Using data in the clear preserves identifiable data so it is possible to use additional techniques to match records with spelling differences and typos.

For example, for the GUiE Wave-1 data, Soundex and distance functions were additional techniques used to match records. Soundex transforms a written name to a code representing their pronunciation in English (for example, "Amy" coded as "A500"). Distance functions provide numerical "distance" between words based on spelling and are used in data linkage to permit a match even when there are minor differences.

For example, an exact comparison between the names "Catherine" and "Katherine" would result in a match status of "false", despite the spellings being almost identical. However, a comparison allowing for a string distance of one would result in a match status of "true". It was possible to use these two techniques as the data were held in the clear. More details regarding these two techniques can be found in our Developing standard tools for data linkage: February 2021 methodology.

The remainder of this section will describe the linkage of GUiE Wave-1 and the joining of Wave-2 data, which took place in the DAP processing environment using data in the clear.

Data preparation

The pre-processing of data for the Wave-1 GUiE linkage included geo-referencing, variable standardisation, and matchkey creation.

Geo-referencing involves referencing data to a specific and fixed point, using a geographic classification and a grid of reference.

Through variable standardisation, all variables are placed on the same scale to allow for comparisons. For example, if an individual's forename is recorded as "Anne-Marie" on one dataset, but as "Annemarie" on the other, the standardisation removes non-alphabet characters and capitalises them. The name will appear as "ANNEMARIE" on both datasets and this forename will link with a 100% field match. Standardisation of variables includes cleaning linkage variables on all data sources to improve linkage rates. Additional processing is completed on the linkage variables to build matchkeys.

Creating matchkeys

The feasibility AEDE and the 2011 Census do not contain a single common identifier that could be used to easily link corresponding records from one dataset to the other. An example of a common identifier might include a unique number for each individual that is contained within both datasets.

Instead, a series of matchkeys containing different combinations of personal identifiable information (PII) were used to link the feasibility AEDE to the 2011 Census. For example, forename initial, surname, date of birth, gender and home postcode may be combined. Any combination would be expected to retain a high level of uniqueness for each member of the population.

Note that for the 2011 Census, imputation was carried out where date of birth had not been provided. Imputation is carried out in a way that preserves the underlying data; other variables or characteristics were used to impute the age of for an individual. For example, if a record referred to an individual being at school, an age range of 6 to 18 years could be inferred. Within our analysis, this enables us to exclude records from the linked dataset based on age.

To help reduce the likelihood of missed matches, 91 matchkeys were created and used. Each matchkey is designed to gradually allow for small amounts of error within the identifier variables, such as difference in name spellings or missing information on the child's gender.

The first matchkey describes an exact match and is of the highest linkage quality. All records matched under this first matchkey are exact matches (no deviation between matched records) for all matchkey elements, meaning they are of the highest linkage quality.

Subsequent matchkeys operate at a lower level of match quality, allowing for small discrepancies between records (for example the missingness or the removal or incorrect reporting of certain identifier variables). For example, the second matchkey allows for errors in surname, while the fourth matchkey does not match on the variable gender. Applying matchkeys in order of strength means that the best quality links are formed earlier.

The 10 strongest matchkeys used for linkage in order of strength were:

1) full forename, full surname, full date of birth, full home postcode, and gender

2) full forename, surname as an alphabetically sorted string, full date of birth, full postcode, and gender

3) forename as an alphabetically sorted string, full surname, full date of birth, full home postcode, and gender

4) full forename, full surname, full date of birth, and full home postcode

5) full forename, surname as an alphabetically sorted string, full date of birth, and full home postcode

6) forename as an alphabetically sorted string, full surname, full date of birth, and full home postcode

7) full forename, full surname, first six characters of date of birth, full home postcode, and gender

8) full forename, surname as an alphabetically sorted string, first six characters of date of birth, full home postcode, and gender

9) forename as an alphabetically sorted string, full surname, first six characters of date of birth, full home postcode, and gender

10) full forename, full surname, first six characters of date of birth, and full home postcode

Three of the 91 matchkeys used in the linkage of the 2011 Census to the feasibility AEDE did not include date of birth. Records matched using these matchkeys are matched on variables other than age, such as home postcode. Therefore, there are a small number of records outside of the expected age range (aged 10 to 25 years at the start of the 2011 to 2012 academic year). Some of these records are likely to be incorrect matches between the 2011 Census and the feasibility AEDE, however others will be correct matches.

Matchkeys were retained only if they were above the uniqueness cut-off threshold. Uniqueness refers to one-to-one matches, where a record from dataset A is matched to just one record from dataset B. For deterministic linkage, this threshold was set at 95%; at least 95% of matches for each retained matchkey were unique. For probabilistic linkage, the threshold was set at a unique count of 15; any probabilistic link made with less than 15 unique matches was ignored.

Deterministic and probabilistic linkage methodologies

Firstly, 2011 Census records were linked to feasibility AEDE records using deterministic, or rule-based, matching. Probabilistic, score-based, linkage was then used to link any unmatched records. Overall, 7,432,115 feasibility AEDE records were linked to the 2011 Census, with 98% (7,255,360) linked by deterministic linkage methods. The remaining 2% (176,758) were linked by probabilistic linkage.

Deterministic linking

The rule-based nature of deterministic linkage means it is often used as an initial step to efficiently remove matches from the pool of total possible matches. This aids further linkage by reducing the population space of remaining records that need matching, and the number of potential links.

Deterministic linkage uses exact matching techniques, where attributes for a given record pair must match according to some rule. The simplest form of deterministic matching is the comparison of all selected attributes in a record pair. When linking the 2011 Census to the feasibility AEDE, the selected attributes for comparison were date of birth, forename, surname, home postcode, and gender. Within the strongest matchkey, only if each of these attributes are identical will the record pair be classed as a match.

Probabilistic linking

Probabilistic linkage was applied to identify additional matches between the 2011 Census and feasibility AEDE records. Within larger datasets, probabilistic methods tend to be more robust against errors.

Probabilistic matching is a score-based method based on statistical theory, which provides a mechanism for setting the weight for each variable. Identical to score-based matching, every record in the first data source is compared with each record in the second, but "blocking" is used as a preliminary step.

"Blocking" is frequently employed within probabilistic linkage; it removes pairs of records that are unlikely to be matched from the pool of possible matches, reducing the number of record comparisons between datasets. For example, a block on name and date of birth would only lead to comparisons of records where both name and date of birth match exactly. Table 1 shows the variables used for this process within census to feasibility AEDE linkage, also called "blocking variables". After "blocking", records between datasets are compared and match-scores are generated.

Table 1: Descriptions of blocks used in Growing Up in England (GUiE) Wave-1 probabilistic linkage and their matchkey uniqueness

Block	Block description	Matchkey uniqueness (%)
92	First six characters of date of birth and full home postcode	85%
93	Full date of birth and first four characters of home postcode	87%
94	First four characters of date of birth and full home postcode	79%
95	First two characters of forename and full home postcode	67%
96	First two characters of surname and full home postcode	67%
97	Full surname and first six characters of date of birth	60%
98	Full surname and first four characters of home postcode	62%
99	First character of forename and full home postcode	67%
100	First character of surname and full home postcode	67%
101	Full forename and first four characters of home postcode	62%

Source: Office for National Statistics - Growing Up in England (GUiE) data

To further identify the strength of the linkage, each matching variable was given a weight (or score) based on its relative use in determining agreement and disagreement.

Agreement is the probability that a variable agrees on two data sources given the pair are a true match. It is a measure of data quality, whereby variables with high data reliability, or freedom from error, are given a higher score.

Disagreement is the probability that the variable agrees on two data sources given that a given pair are a true non-match. It measures the distinguishing power, whereby variables with high distinctiveness are given higher scores. These probabilities were combined to provide a single weight for each matching variable.

As an example, gender can be given a high agreement score but low disagreement score. Two records belonging to the same person are likely to match on the gender variable (a true match) but the limited categories available within this variable mean that two records belonging to different people are also likely to match (a true non-match). On balance, the latter outweighs the former in this example, and so gender is given a low weight.

Table 2 shows the weight for each matching variable used for the GUIE Wave-1 linkage.

Table 2: Description of matching variable and weighting

Description of matching variable	Weighting
Forename	2
Surname	3
Gender	1
Year of birth	1.5
Year and month of birth	0.75
Day of birth	0.75
First two characters of home postcode	1.2
First four characters of home postcode	1.2
Last three characters of home postcode	0.6

Source: Office for National Statistics - Growing Up in England (GUiE) data

Clerical linking

A clerical review on the feasibility AEDE data was conducted to determine the selected threshold above which rates of true and false positives and below which rates of true and false negatives were considered acceptable.

True positives are pairs of records that are correctly linked, and false positives are pairs of records incorrectly linked. True negatives, in contrast, are pairs of records that were correctly identified as having no link. False negatives are pairs of records where the link was incorrectly missed.

From a random sample of 901 records, samples of around 35 matched records each were generated. Each match was clerically classified as "true", "false", or "maybe" and classified samples were appended back into a single file.

Records were classified as "true" where clerical review determined that these linked records belong to the same individual (records identified as true positives). "False" referred to records which were linked but, upon clerical review, were identified as belonging to different individuals (records identified as false positives). Finally, records were classified as "maybe" where clerical review was unable to determine whether these records belonged to the same individual or not.

Following this, the cut-off threshold was determined using aggregate analysis where frequencies of "true", "false", and "maybe" classifications were computed by a 0.01 match score bracket. A threshold was selected so that above the threshold rates of true and false positives and below the threshold rates of true and false negatives were considered acceptable.

Wave-2 join

While Wave-1 feasibility AEDE data were linked to census data, Wave-2 data could be joined to Wave-1 using the person ID present in the education and vulnerability data.

For the Wave-2 join, files were imported and run against the data spine that contains the person ID to extract person IDs. Records with IDs not already found on the person spine were disregarded to then be joined on later. Records containing person IDs in the Wave-1 and Wave-2 data were joined together on the person ID. Next, unjoined records were identified. The unjoined records received a newly generated Administrative Data Research (ADR) ID in a format that differs from IDs created as part of Wave-1. This means that we can determine whether a record has been appended to the person spine that did not match to a Wave-1 record.

5 . Quality assurance and the production of a record duplication flag

Quality checks

Quality checks were conducted to assess the quality of match rates within the feasibility All Education Dataset for England (AEDE). The feasibility AEDE is an administrative data source, meaning data are collected for administrative, not research, purposes.

Where administrative data are used for research purposes, it saves time, resource and lessens respondent burden where survey data collection is used. However, the quality of administrative data is expected to be lower than the quality of survey data or any data collected for specific research purposes. Lower quality data makes the linking of data more difficult, so it is important that data linkage is thoroughly assessed.

Quality checks involved an assessment of match quality within the feasibility AEDE. Precision and recall rates were used to give an indication of match quality. A sample of the feasibility AEDE (900) was taken. The number of false negatives (41), true positives (541), and false positives (108) were calculated using a threshold above which rates of true and false positives and below which rates of true and false negative were considered acceptable. For the feasibility AEDE, 0.75 was the suggested cut-off score.

Precision rate is calculated by dividing the number of true positives by the total number of matches and is an indication of match accuracy. Recall rate is a sensitivity measure, calculated by dividing the number of true positives by the sum of true positive and false negative matches.

The precision rate was 83.4%, while the recall rate was 93.0%. Note that rates were calculated on the assumption that samples classified as "maybe" in the clerical review are assumed to be and therefore classified "false". Not all records originally classified as "maybe" are likely to be false positives, so precision and recall rates are likely to be higher.

Further quality checks uncovered several issues with the core datasets that make up Growing Up in England (GUiE). The Office for National Statistics (ONS) undertook a clerical matching exercise that revealed duplication within the feasibility AEDE at the source; some identifiers (IDs) had multiple different associated records, which prompted further clerical work.

Clerical linkage exercise

A clerical linkage exercise was used to address the duplication observed in the feasibility AEDE data. Firstly, 100,000 ID numbers were ordered and sampling of 100 IDs (every 1000 rows) was used to determine the approximate index range at which the ID numbers begin to duplicate within the data.

Personal information (for example name, date of birth, and postcode) was compared for each record associated with an affected ID number, to determine whether the information belonged to an individual or multiple people. Duplicates were recorded where multiple individuals were assigned the same ID. Duplicates were given the decision "true". Where an instance of multiple records belonged to the same individual, these were given the decision "false". In cases where a decision could not be confidently made, these were given the decision "maybe".

Having identified the range, a further 33 ID numbers were sampled to determine to a better degree the approximate location of the threshold. The breakdown of "true", "false", and "maybe" decisions is shown in Table 3.

Table 3: Breakdown of clerical linking decisions

Decision	Count
TRUE	39
FALSE	84
MAYBE	10
Total	133

Source: Office for National Statistics - Growing Up in England (GUiE) data

The approximate threshold was identified to lie around the 23,300 case, at which point ID numbers have an index score of 40. Overall, the higher the index score, the greater likelihood that there would be an ID number duplication.

As part of the clerical matching exercise, the following data quality issues were identified.

Variation in names

There were records belonging to the same individual where names had been misspelled or recorded differently at different times (for example nicknames, name changes, and transposition of forename and surname).

In some instances, different people, evaluated as such because of differing genders, forenames, and dates of birth, had been assigned the same middle name and surname. It was deemed unlikely that these individuals were siblings because of the implausibility that so many siblings shared the same middle names. Instead, these findings suggest an underlying quality issue with the data and/or linkage.

Date of birth error

There were some minor issues in recorded date of birth for some individuals. For example, one digit in a date of birth changed between records for the same individual, suggesting an input error at source. In other cases where date of birth differed significantly alongside differences in other variables, these records were deemed to be true duplicates.

Date of birth was imputed in cases where date of birth was not recorded for an individual. The imputation was based on surrounding information; if the respondent was at school age, their date of birth may be imputed to an age between 5 and 16 years, as these are the typical ages that a child attends school.

Gender error

In multiple cases, feasibility AEDE records for the same person were repeated with the opposite gender assigned. However, we acknowledge that gender is not necessarily stable over time. As such, this variable may not be the strongest predictor of record duplication.

Postcode repetition

In multiple cases assigned true duplicate status, postcodes followed the same pattern across different sets of duplicates. This again suggests an underlying quality issue with the data or linkage because of the unlikelihood that so many people moved to the same postcodes.

This exploratory work identified several variables considered to be predictors of duplicate records. Subsequently, a regression analysis was run to derive an improved predictor of duplicate records.

Regression analysis to produce record duplication flag

A sample of 165 records were taken from an area with a higher concentration of duplicate records compared with the overall dataset. From previous clerical work, duplicates were classified as either "true", "maybe", or "false". After determining that regression assumptions were met, a logistic regression was used to analyse these data.

Logistic multiple regressions were run on the data to determine the best predictor for duplication. Three investigations were run, each handling duplicates classified as "maybe" in a different way, meaning they were either:

- removed from the analysis
- classified as true
- classified as false

Regressions were run on all variables selected for the model then non-significant predictors were removed and the regressions run again. The testing process involved including and excluding categorical variables to observe the effect on the models. The equation that produced the highest pseudo R-squared was taken as the best predictor equation, as this equation explained the greatest percentage of variance. Within this equation, "maybe" records were removed from the analysis and categorical variables were excluded. Odds ratios and confidence intervals were produced for this logistic regression (Table 8).

Table 4: Regression model with “maybe” records removed

	Coefficient	Standard error	Z-score	P-value	Lower 95% confidence interval	Upper 95% confidence interval
Intercept	-7.49	1.98	-3.79	0.00	-11.36	-3.61
Boolean indicating whether there is a record for postcode	0.17	0.84	0.20	0.84	-1.48	1.81
Boolean indicating whether ADR ID has multiple records for postcode variable	-2.62	0.97	-2.70	0.01	-4.51	-0.72
Total names associated with the same ADR ID	0.20	0.27	0.74	0.46	-0.32	0.72
Number of unique surnames associated with the same ADR ID	1.23	0.41	3.03	0.00	0.43	2.03
Number of unique forenames associated with the same ADR ID	0.71	0.47	1.51	0.13	-0.21	1.63
Number of unique dates of birth associated with the same ADR ID	1.13	0.60	1.89	0.06	-0.04	2.30
Number of males associated with the same ADR ID	0.10	0.03	3.05	0.00	0.04	0.17
Number of females associated with the same ADR ID	0.082	0.04	2.31	0.02	0.01	0.15

Source: Office for National Statistics - Growing Up in England (GUiE) data

Table 5: Summary of regression model data

Number of observations	139
Degrees of freedom (residuals)	130
Degrees of freedom (model)	8

Source: Office for National Statistics - Growing Up in England (GUiE) data

Table 6: Regression model with “maybe” records and non-significant predictors removed

	Coefficient	Standard error	Z-score	P- value	Lower 95% confidence interval	Upper 95% confidence interval
Intercept	-3.93	0.86	-4.55	0.00	-5.62	-2.24
Boolean indicating whether multiple postcodes are assigned to a record	-2.68	0.86	-3.12	0.00	-4.36	-1.00
Number of unique surnames associated with the same ADR ID	1.24	0.34	3.67	0.00	0.58	1.91
Number of males associated with the same ADR ID	0.12	0.03	4.19	0.00	0.07	0.18
Number of females associated with the same ADR ID	0.09	0.03	3.39	0.00	0.04	0.15

Source: Office for National Statistics - Growing Up in England (GUiE) data

Table 7: Summary of regression model data

Number of observations	139
Degrees of freedom (residuals)	134
Degrees of freedom (model)	4

Source: Office for National Statistics - Growing Up in England (GUiE) data

Table 8: Odds ratios and confidence intervals

	Lower 95% confidence interval	Upper 95% confidence interval	Odds ratio
Intercept	0.00	0.11	0.02
Boolean indicating whether multiple postcodes are assigned to a record	0.01	0.37	0.07
Number of surnames associated with the same ADR ID	1.78	6.72	3.46
Number of males associated with the same ADR ID	1.07	1.20	1.13
Number of females associated with the same ADR ID	1.04	1.16	1.10

Source: Office for National Statistics - Growing Up in England (GUiE) data

The regression equation produced following this analysis was in agreement with 91% of the clerical decisions (excluding records classified as "maybe") with the sample of feasibility AEDE records. Since the sample focussed on areas of uncertainty, it is likely that the true rate of agreement will be higher than this for the entire feasibility AEDE population. This equation informed the production of a record duplication flag for researchers to use, when accessing the GUiE dataset within the Secure Research Service (SRS).

In summary, following the issues highlighted after data quality assessment exercises were undertaken, we have produced a record duplication flag and data quality manual available to users of GUiE data in the SRS. Researchers within and external to the ONS have used various components of GUiE in their research. We have incorporated feedback and are continually working to update and improve the existing metadata and documentation surrounding the GUiE data.

6 . Analysis of Wave-1 and Wave-2 of Growing Up in England (GUiE)

This methodology links to data tables that show the types of analysis that are possible using the Growing Up in England (GUiE) dataset. These data tables show new insights that can be gained by linking education and vulnerability measures data to census data. They also provide researchers with an indication of the sample sizes they can expect, when using the GUiE data.

Data preparation for analysis of the Growing Up in England (GUiE) dataset

Datasets were joined using a unique identifier (ID) that identifies records across data sources and academic year where possible. Duplicate records were dropped in terms of the unique ID and academic year where applicable, so that records were dropped in cases where there were multiple records per academic year. This was done to preserve the numbers of unique students in the data as much as possible; there would only be one record of a given student per academic year.

A record duplication flag was produced to address record duplication observed in the feasibility All Education Dataset for England (AEDE).

For more information on the production of the record duplication flag, see Section 5: Quality assurance and the production of a record duplication flag.

Instances of true duplication, whereby multiple records on the same unique ID were deemed to relate to different people, were removed from the dataset using this flag. In these cases, the entire record of the individual was removed from the dataset. Of the 27,043,210 records in the person spine, 5,745,700 were deemed true duplicates.

For a small number of records, age recorded at the time of the 2011 Census was outside of the expected age. This is as an anticipated result of the linkage methods used. Records with ages outside the expected age range were excluded; however, an administrative error resulted in the inclusion of records aged between 10 and 29 years at the time of the 2011 Census. However, the feasibility AEDE cohort is aged 9 to 25 years at the time of the 2011 Census. There are some children aged nine years at the time of the 2011 Census (27 March 2011), who turn 10 years before the end of the academic year (31 August 2011). These children are not represented within our analysis because it was not possible to determine which nine-year-old children turned 10 before 31 August 2011. Only individuals aged between 10 and 25 years at the time of the 2011 Census – 7,100,509 records – should have been retained for this analysis. This meant that 8,498 records of individuals aged between 26 and 29 years at the time of the 2011 Census that should have been excluded based on our exclusion criteria, were included in this analysis.

As such, where permitted, age was restricted to those aged 10 to 29 years using age at the time of the 2011 Census. Where possible, tables document missing values. These are where the linked records do not contain a record for a particular variable.

Tables produced as part of the GUiE analysis

Tables produced as part of this analysis are within our accompanying datasets.

Our five datasets contain several tables within separate tabs of an Excel workbook. Here these groups of tables are listed, with a brief description of the analysis conducted.

Personal characteristics and educational attainment

Personal characteristics and educational attainment tables are available within this accompanying dataset.

These tables show information on the population sizes of children broken down by a range of personal characteristics and educational attainment.

A number of individuals had to be excluded from the analysis because of issues that rose post-linkage. These post-linkage issues meant it was not possible to combine children's educational attainment information with information on personal characteristics collected from the census.

Records assigned the same ID that were determined to relate to multiple individuals were excluded from this analysis. Records for which there were more than one entry per academic year were also excluded in each dataset that required one-to-one matching. This step was necessary to optimally preserve the true number of records. However, in doing this, any records with long-format variables would potentially be dropped. Long-format variables are variables where individuals' data are contained within multiple rows where there is more than one entry for that individual.

Individuals aged under 10 years or aged 29 years and over were excluded.

For the 2011 Census standalone dataset, the total population of cohort children and young people fell from approximately 7.4 million to approximately 7.1 million.

A linked National Pupil Database (NPD) dataset was created, which included:

- NPD Attributes, obtained from the spring term English School Census (ESC), pupil referral units, and alternative provision censuses
- NPD Attainment, obtained from the Young Person's Matched Administrative Dataset (YPMAD)
- person spine
- 2011 Census

For the linked NPD dataset, the total population fell from approximately 5.6 million to approximately 3.3 million children in 2010 to 2011, and approximately 5.4 million to approximately 3.2 million children in 2014 to 2015.

A linked Individualised Learner Record (ILR) dataset was also created, which included:

- ILR Aims
- ILR Learner
- person spine

For the linked ILR dataset, the total population fell from approximately 2.1 million to approximately 1.6 million children in 2010 to 2011, and approximately 2.5 million to approximately 1.9 million children in 2014 to 2015.

Household characteristics and educational attainment

Household characteristics and educational attainment tables are available within this accompanying dataset.

These tables show educational attainment of children by household composition. Analysis was carried out on household-level characteristics. These datasets were also deduplicated and age restricted.

General breakdowns

General breakdown tables are available within this accompanying dataset.

These tables show educational attainment by geographic region. Geographical information was correct at the time of the 2011 Census.

To produce the counts of students across the ILR and NPD, duplicate records were removed according to the record duplication flag for both datasets.

See Section 5: Quality assurance and the production of a record duplication flag for more information.

Where possible, age was restricted to those who were aged between 10 and 29 at the time of the 2011 Census.

ILR tables

ILR tables are available within this accompanying dataset.

ILR data record attainment of children in further education and work-based learning in England. The tables provide information on the types of provision and counts of how many are in the main categories.

Tables were produced using ILR data merged with the person spine to remove instances of record duplication. In some cases, this step was not taken to minimise disruption of the true number of individuals in the data. In these cases, a note is added stating that ILR data is not merged with the person spine. Technical limitations at the time of producing ILR tables resulted in 2011 Census data being excluded from the merged dataset in some cases. This will be notated in the corresponding tables.

Coverage of the ILR varies. This is reflected by the coverage of the tables presented. The tables are produced either to demonstrate the full range of linked data across the available academic years, or to show a snapshot of attainment in the 2010 to 2011 and 2014 to 2015 academic years.

Vulnerable groups and educational attainment

Vulnerable groups and educational attainment tables are available within this accompanying dataset.

There are a number of characteristics used by the Office of the Children's Commissioner for England (CCO) to identify vulnerable children. The characteristics of vulnerable children that are captured within the Wave-1 dataset are children from minority ethnic backgrounds, young carers, and children and young people living in lone-parent families.

The Wave-2 data capture information on Special Educational Needs, Free School Meals, disability, looked after status and primary need. Information on vulnerability characteristics was sourced from the Census, Absences, Exclusions, English School Census (ESC), Children in Need (CIN) and Children Looked After (CLA) datasets.

For the CIN data, each child had one record per disability in a given academic year. For example, if a child had two disabilities, there would be two records of disability for a given academic year, as opposed to a single record containing the two disabilities. When linking datasets for this analysis, duplicates in terms of ID and academic year were not removed from the sample in order to preserve the instances of different disabilities within the data. As such, the total numbers of students in the disability outputs may appear larger than the true number of students.

Additional notes on our analysis of Growing Up in England (GUiE) Wave-1 and Wave-2

This analysis is similar to the analysis we conducted within our Proof of Concept (PoC) study, and we aimed to reproduce tables that were originally created using the PoC. In our previous PoC, we were able to derive full family composition for households containing up to six members. However, this analysis is not possible as only census records matched to the feasibility AEDE are contained within the data. Therefore, we are unable to explore the differential impact of a parent's demographic and socioeconomic characteristics on educational outcomes as before. This census extract particularly affects the tables produced within "Household characteristics and educational attainment".

However, using Wave-2 vulnerability data that was not available within the PoC, we are able to produce additional tables which explore vulnerable groups and educational attainment.

Educational attainment records from 2001 to 2002 to 2014 to 2015 are linked to 2011 Census data. Accordingly, personal, demographic, and geographic information, though correct at the time of the census (27 March 2011), may be incorrect at the time the educational data were obtained.

This can help to explain the presence of no-children-households according to the 2011 Census measure of household composition. The oldest individuals contained within the first year of GUiE data are aged 24 or 25 years at the time of the 2011 Census, so they may be recorded as living alone, or with a partner. Alternatively, households with no children can refer to households with non-dependent children. A child is considered non-dependent if they are aged 16 to 18 years and not in full-time education, and have no spouse, partner or child living in the household.

Information as recorded in the 2011 Census may also have been provided by the parent, guardian, or other household member rather than by the child self-identifying.

7 . Future developments

At present, some terms of English School Census (ESC) are missing from the Wave-2 vulnerability data. These include spring 2014 to 2015, summer 2014 to 2015, and autumn 2010 to 2011 to 2012 to 2013. Wave-2 data are only available for the academic years 2010 to 2011 to 2014 to 2015, whereas Wave-1 data are available for academic years 2001 to 2002 to 2014 to 2015. We aim to deliver a third wave of data into the Secure Research Service (SRS), which will contain, where available, these additional years of vulnerability data and terms of ESC.

There are two versions of the 2011 Census: person and household. Both include person-level information. However, the household version additionally includes IDs of each household member and their corresponding census information. This enables exploration into the relationships between different household members and how these interact with educational outcomes. Future waves of Growing Up in England (GUiE) could link the household census to the feasibility All Education Dataset for England (AEDE) to increase understanding of the role of familial relationships and educational outcomes.

Some variables, such as ethnicity, are present in multiple GUiE source datasets. We are working to assess whether these variables are congruent across datasets.

Discussions are underway to assess the possibilities of a new build of GUiE using the Census 2021 linked to the Office for National Statistics (ONS) built All Education Framework for England, after a pause and reflect period.

8 . Accessibility of the data

For Growing Up in England (GUiE), permission was obtained from the Department for Education (DfE) and the Office for National Statistics (ONS) census team to link their data for research purposes.

The Statistics and Registration Service Act (SRSA) 2007 and Digital Economy Act (DEA) 2017 provided the legal gateways for the data to be accessed through the ONS. As a data processor, we cleaned, linked and de-identified this data. It was then made available for accredited researchers to access on the Secure Research Service (SRS), in line with the Five Safes framework.

The GUiE data can be accessed by all accredited researchers for research purposes through the ONS SRS.

The research strand of the DEA 2017 enables researchers to access data for research purposes through accredited processors. The SRS was accredited to store and provide accredited researchers with access to de-identified data from 12 August 2019. Researchers can apply for accreditation through the Research Accreditation Service (RAS). See further information about the Secure Research Service (SRS) and accessing data using the Research Accreditation Framework.

To request access to data in the SRS, researchers must also submit a research project application. All DEA research project applications will be considered by the UK Statistics Authority (The Authority) and accredited using a framework agreed and overseen by the Research Accreditation Panel (RAP). The RAP was established by The Authority to independently consider applications for research, researcher, or processor accreditation. See further information about the Research Accreditation Panel.

All applications for project and researcher accreditation made under the DEA are subject to the Research Code of Practice and Accreditation Criteria.

A complete record of accredited researchers and accredited projects is published on The Authority's website to ensure transparency of access to research data.

9 . Glossary

Data linkage

The act of bringing two or more datasets from different sources together, creating associations between the data. Data linkage can provide new statistical insights not possible with information from a single source.

Data processing

Data processing is the method applied to convert data into a format that can be interpreted, analysed, and used for a variety of purposes.

Data quality

An essential characteristic that determines the reliability of data for making decisions. High-quality data are complete, accurate, available, and timely.

De-identified

De-identified data do not contain any personal identifiable information, such as name, address, or postcodes. The identifiers are removed from the records before de-identified microdata are securely transferred to the Trusted Research Environment (TRE; a secure research system where access is controlled).

Hashing

The practice of using an algorithm to map data of any size to a fixed length, for example the hashed value. This string of characters is non-disclosive. When the same algorithm is applied to the same string of data the same hash value is produced. This makes it possible to link the data without disclosing the identifiable information. Hashed data cannot be "un-hashed"; you cannot pass it back through an algorithm to get the starting value. However, this does not mean it cannot be undone, for example hacking or frequency analysis. Techniques such as using salts mitigate against the risk of hacking.

Pseudonymised

The processing of personal data in a way that the personal data can no longer be attributed to a specific entity without the use of additional information. Additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable entity.

10 . Related links

Educational attainment and household composition, feasibility research and methodology
Methodology | Released 30 July 2020
Research Outputs of the feasibility study that links the All Education Dataset for England (AEDE) to the 2011 Census. The new de-identified Growing up in England (GUiE) dataset will enable research into the link between family household composition and educational attainment.

Childhood vulnerability in England 2017
Report | Released 4 July 2017
Report conducted by the Children's Commissioner for England, which revealed the scale of childhood vulnerability in England.

Secure Research Service (SRS) Metadata Catalogue
Webpage | Released February 2022, updated as and when data become available within the SRS
Metadata catalogue containing further information about the data available within the Growing Up in England (GUiE) dataset.

11 . Cite this methodology

Office for National Statistics (ONS), released 18 October 2022, ONS website, methodology, Growing Up in England (GUiE)