Ministry of Justice – Department for Education linked dataset

Feasibility of evaluating early interventions for violence prevention:

Data quality report

July 2022

Economic and Social Research Council

ADR UK

A primer on record linkage

Record linkage, a process in which an individual's records held in separate datasets are linked together, can be used to create novel and powerful data sources to drive evidence-based policy and practice. In England, where there is no universal citizen ID number, linkage between departmental datasets is a non-trivial undertaking. Data flows for linkage must have a legal and regulatory basis and need to address substantive public concerns. Typically, this results in access to identifiers being restricted to a minimum number of individuals and the linkage process happening separately from the subsequent reuse of the data. Without rigorous assessment of the processes and supporting documentation, analysts can be blinded to data quality issues arising from the linkage.

Linkage attempts to reconcile and match sets of personal identifiers, which may be subject to recording error and can change over time, thereby impacting on the accuracy of this process. To address this, those performing the linkage conduct pre-processing to harmonise identifiers (for example, reconciling "Jon Smith", "Jonathon Smith" and "Jon Smiht" to "JONSMITH"). Different linkage methods can also be used: deterministic linkage seeks to identify exact matches whilst probabilistic linkage aims to estimate the probability that two records are from the same individual even when variation exists across the identifiers. Fuzzy linkage within a deterministic approach seeks to account for variation by conducting multiple passes through the records and each time adjusting which identifiers (or parts of identifiers) are used. Clerical review can be used to make decisions on marginal matches or to determine probability score thresholds used to classify whether a match is deemed to be true or false.

Linkage error occurs in the form of false matches, where two records from different individuals are incorrectly matched, and missed matches, where records from the same individual are not linked. Error may be compounded when the source data are products of internal linkages with their own errors. If random, false matches generally weaken associations between variables from the linked datasets whereas missed matches tend to reduce precision (although can also introduce bias). However, linkage error has been shown to disproportionately affect more vulnerable or marginalised groups (i.e. not occur randomly), resulting in the linked data not being representative of the population of interest and perhaps even excluding particular subgroups. This differential error can lead to bias; the extent of this will depend on the specific research question being addressed using the data. Typically, a linkage process will aim to minimise false matches whilst also minimising missed matches. However, without a gold standard dataset in which the true match status is known, error rates cannot be determined.

The overall objective and metric of success of a linkage exercise should not be a headline match rate, but rather heterogeneity in the linked sample and a rigorously documented process, complete with a match quality indicator. Further, because the impact of linkage error in terms of bias will be very project-specific, data users should have some means of comparing matched and unmatched individuals. All this should enable informed inferences to be drawn from the data.

Executive summary

Background

In 2020, supported by the ADR UK-funded 'Data First' data linkage programme led by the Ministry of Justice (MoJ, 2021), a new database was established linking data from the criminal justice system with Department for Education (DfE) data from the education and social care systems. To link the datasets, the MoJ provided the DfE with identifiers of all offenders born from 31st August 1985 with at least one caution or conviction from 2000 or later. The DfE linked these identifiers to identifiers recorded in in the National Pupil Database (NPD) and other DfE datasets.

The UK does not have a national registry with a unique personal identification number, so cross-departmental data sharing requires novel data linkage processes involving considerable reconciliation of information and subsequent determination of the linkages. Linkage error, which arises as a result of either false matches or missed matches, can introduce biases in research findings if particular subgroups in the population are less likely to be matched or more likely to be mis-matched. In order for users to assess the likelihood of this, it is important for them to understand the linkage methods and to be able to assess whether unmatched individuals differ systematically from matched individuals.

The project

This project was commissioned by ADR UK, on behalf of the Home Office, to assess the feasibility of using the MoJ-DfE linked dataset for evaluating early interventions for violence prevention. The project had two stages: 1) to assess the reliability of the data included in the linked dataset; and 2) to investigate whether the linked dataset could be used to generate matched control groups for evaluation purposes. In this report – relating to stage one of the project – we describe the data sources, the linkage process and a range of quality assessments carried out on a subset of variables from a subset of datasets included in the data share. Our options for evaluating linkage quality were restricted as our assessment was conducted after the linkage had been designed and conducted and we had no access to the personal identifiers used in this process or to the input files. We therefore concentrated on documenting the linkage process – through retrospective collection of MoJ and DfE written and verbal information – against the GUILD linkage reporting framework.

Key findings and recommendations

The final dataset we evaluated included just over 15 million individuals born between September 1985 and August 2007 with education records, linked to 1,842,478 individuals with criminal justice outcomes. As a result of coverage of the key DfE datasets, individuals who attended independent schools or were home-educated throughout their schooling will be effectively excluded[1] from the linked dataset. In addition, individuals attending independent schools or home-educated for the majority of their schooling as well as those who were educated outside England for the majority of their schooling, would have limited (or no) education data present in the linked dataset.

Finding 1: In the linked data, completeness levels were high and levels of inconsistency were generally low. Completeness in variables in the NPD datasets generally improved over time.

In any large administrative dataset, some degree of incompleteness and inconsistency is inevitable. However, we found low levels of incompleteness and minimal inconsistency across most core measures; the main exceptions to this were DfE-recorded ethnicity and Key Stage 2 and 3 attainment data. Just over 6% of individuals had no ethnicity recorded or ethnicity classified as refused or not yet obtained. A further 14% had more than one minor ethnic group recorded. In terms of attainment, up to 8% (Key Stage 2) and up to 10% (Key Stage 3) of records contained invalid data; these percentages were higher among those individuals with offending data.

Finding 2: 70% of individuals with MoJ identifiers were matched to an individual in the DfE data sources. Because of the differences in geographical coverage between the MoJ and DfE data sources, some individuals would not be expected to match.

Of the 2,631,842 individuals in the MoJ's Police National Computer (PNC) and Home Office Court Appearance Statistics (HOCAS) source datasets, 27% were unmatched. An additional 3% were removed (where the linkage resulted in matches to multiple individuals) leaving 1,842,478 individuals with criminal justice outcomes (70% of the original 2,631,842 individuals with at least one recorded caution or conviction from 2000 or later). Future shares are intended to only use the PNC source dataset. The final match rate for the PNC data was 77% (1,512,106 of 1,953,599 individuals). Recent analyses by the MoJ addressed the variation in geographical coverage between the two datasets (given the likelihood that a percentage of those with offending data were educated outside of England, and therefore would not have a record in the NPD). For individuals with a main home postcode in Wales, 82% did not match, whereas this proportion was only 8% for English postcodes (MoJ 2021b).

[1] These individuals would appear in the dataset – with attainment data - if they sat GCSEs or other regulated qualifications but would have no other data.

Match rate is distinct from match quality, where the rates of false matches and missed matches are the primary objects of interest.

Finding 3: The match quality indicators are useful measures to include in the dataset

Just under two thirds (64%) of linked individuals matched exactly on all identifiers. The match quality indicators suggest that the linkage was more certain when the MoJ identifiers came from the PNC (the PNC is law enforcement agency recorded crime) rather than the HOCAS data (this includes data passed directly to court by the relevant agencies, for example motoring and credit or license default, and other non-recordable offences) and when the DfE identifiers came from the school census data – rather than the Individual Learner Record (ILR), attainment or Higher Education Statistics Agency (HESA) sources. The match quality indicators are a useful addition to the core dataset, as researchers can use this information to carry out sensitivity analyses relevant to their study (for example, restricting on linkage certainty, with a view to minimising false positives).

Recommendation 1: Future cross-departmental data linkage protocols could potentially be refined following consultation with a wider range of linkage experts and all processing stages should be fully documented.

There were key linkage stages that involved manual processing with limited documentation. Therefore, the potential for the linkage protocol to introduce quality issues and potential bias cannot be fully described. Full transparency of all processes would increase clarity in this area and may facilitate future efficiencies in user support and linkage design. It would also enable researchers to fully assess the likely impact of data linkage quality on their project.

Linkage expertise exists within the departments who act as custodians of the source data, and also in the end-user community in academia. Involving a small group to assess and advise on the marginal gains from refining protocols over time would positively contribute to the accuracy of the linked data.

Recommendation 2: Consider including a table within the metadata and additional detail in the access process to describe the pseudonymised data on unmatched individuals, along with their key attributes.

Providing individual-level attribute data on unmatched cases allows user assessments of potential biases tailored to the specific purpose of their research. The MoJ has reported aggregate counts, comparing unmatched and removed records with matched and retained records by key attributes. This helps identify patterns in matching from which inferences can

be drawn. For example, the increased likelihood of matching for younger offenders, possibly because their identifiers are less likely to have changed. Making unmatched individuals and their attributes (age, gender, ethnicity etc.) visible in the dataset by including details of these data in the metadata would encourage researchers to apply for these data and make a full assessment of the degree to which linkage error might have impacted their findings; this will inevitably vary depending on the research questions being addressed.

Contents

1 Introduction

In 2020, the Ministry of Justice (MoJ) and the Department for Education (DfE) carried out a data share, linking data from the criminal justice system, including police, prison, and court records, with data from the education and social care systems. The MoJ provided the DfE with identifiers of all offenders born on or after 31st August 1985 with at least one caution or conviction from 2000 onwards. The DfE linked these identifiers to identifiers of individuals appearing in the National Pupil Database (NPD) and other DfE datasets.

The dataset includes individuals captured at least once in the education record across all the datasets included in the linked data (for example, those attending a state school in England during their reception year but who then moved outside England for the remainder of their schooling). Characteristics of those in alternative provision, or at an independent specialist provider (including childminders) are in the data. Attainment data are returned for all individuals in state-funded schools up to Key Stage 5; the dataset also includes attainment data for pupils in non-maintained special schools, sixth form and further education colleges and, for Key Stage 4 and 5, those attending independent schools.

This report is based on the first stage of a wider project – commissioned by ADR UK on behalf of the Home Office – carried out to assess the feasibility of using the MoJ-DfE linked dataset for evaluating early interventions for violence prevention. We applied and were granted access to the subset of datasets and variables from the linked data needed to carry out the feasibility study (the list of datasets and variables we requested are given in Appendix B). The dataset we were granted access to via the Office for National Statistics Secure Research Server (ONS SRS) included education data up to the academic year 2017/2018 where available.

The aim of this first stage of the project was to assess the reliability of the data included in the linked DfE-MoJ dataset. This report has two parts. Part 1 describes the source data and data linkage, and has four sections. In Section 2 we give an overview of the resource, detailing the datasets included and their coverage. In Sections 3 and 4 we use information provided to us by the MoJ and DfE to report – retrospectively – on the linkage process and the match quality.

Analysts have a range of methods for dealing with data quality issues, including linkage error, but cannot do this in the absence of other information. The answer for data owners is to provide transparency; for linkage this can be done by identifying, documenting, describing and quantifying each step where individuals from both data sources have missing data, are matched, unmatched or dropped.

The GUidance for Information about Linking Data sets (GUILD) offers clarity on what information is required (for research involving linked datasets) to minimise the introduction of error, reduce bias, and improve the validity of subsequent analysis and the interpretation of results. Note that this is focused purely on linkage quality not on other data quality issues. It breaks down the data cycle of a project involving linked data into four stages:

1. Data provision - the generation, processing and quality control of the source data for linkage
2. Data linkage - bringing together records belonging to the same individual, place or organisation
3. Analyses of the linked data - taking account of linkage error
4. Reporting the results of analyses of linked data

To meet the objective of transparency, we address the detail in points 1 and 2 by describing the source datasets, pointing toward available supporting documentation, and describing the linkage process. We were provided with documentation where it was available and data on matched (but not unmatched) cases, including indicators of match quality.

In Part 2, we report on quality checks carried out on the subset of datasets and variables included in our extract of the linked dataset, focusing in particular on completeness, consistency, and uniqueness. This part has six sections. Section 5 gives an overview of this part of the work and Sections 6 to 10 focus on different aspects of the data.

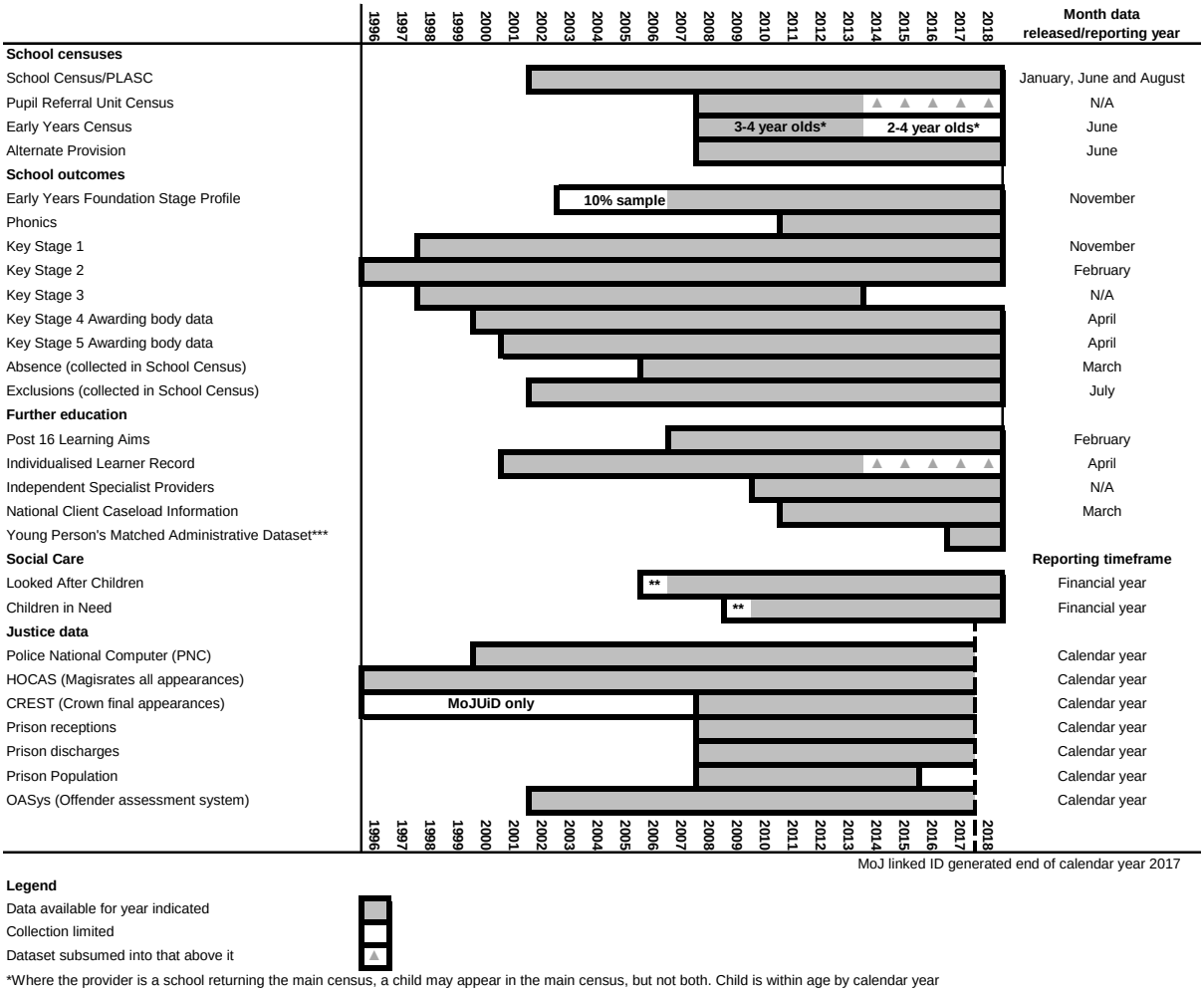
At the end of the report (Section 11) we provide an overview of the findings from parts one and two of this stage of the project and discuss the key limitations and caveats of this work.

PART ONE: Source data and data linkage

2 Source data

The source data includes a number of different education, social care and justice datasets, which cover different periods of time (Figure 2.1). In this section we briefly describe each of these datasets.

Figure 2.1: Overview of the datasets included in the MoJ-DfE data share



Legend

Data available for year indicated

Collection limited

Dataset subsumed into that above it

*Where the provider is a school returning the main census, a child may appear in the main census, but not both. Child is within age by calendar year

**Social care data are reported on 31 March every year. It is not possible to link data before the years shown

*** formerly Level 2 and 3 at 19 indicators

With thanks to: Jay, M.A., Mc Grath-Lone, L., Gilbert, R., 2019. Data Resource: the National Pupil Database (NPD). *International Journal for Population Data Science* . 4(01) 08.

2.1 Department for Education datasets

2.1.1 National Pupil Database (NPD)

The NPD is an administrative data resource curated by the DfE and used for funding purposes, school performance assessments, policy making, and research (Jay et al. 2019). The resource comprehensively details a child's journey through the state-maintained education system from early year providers (including childminders), infant and junior school (Key Stages 1 and 2), through secondary school (Key Stages 3, 4 & 5), to school exit.

The NPD also contains two sources of administrative social care data, both of which have been documented previously: Children Looked After (CLA) return (McGrath-Lone et al. 2016) and Children in Need (CiN) (Emmott et al. 2019). Social care data are submitted by the local authority when a child is referred to social services for a needs assessment (CiN), or when the child has an episode of out-of-home care arranged by the local authority (CLA).

The NPD is comprised of individual statutory collections that are returned, either yearly or termly, by different organisations (including local authorities, exam awarding bodies, state-funded primary and secondary schools, and further education providers). These collections are linked together to form the NPD. Technical and business documentation, detailing the entry rules, validation checks and processing of the data sources are publicly available and held by the DfE (DfE, 2021) or stored within the National Archives (The National Archives, 2021). They are produced at least yearly, with version control, and are a valuable source of metadata, as the onus for data accuracy is placed on those submitting the data, and rules change over time. School census data (see below) are complete at the point of release, whilst other components are 'living', and subject to updating when new information is submitted.

2.1.1.1 School census data

School census (pupil-level) (2001/02 onwards)

The school census is a pupil-level termly statutory return for all state-maintained settings[2] in England. Pupil referral units were included from academic year 2012/13. The school census includes enrolment data, basic demographic data such as age, gender, and ethnicity as well as information on the pupil's first language, special educational needs (SEN), eligibility for free school meals (FSM) and other indicators. It also has codes indicating the school and local authority. The school census also captures data on absence and exclusions; these data are supplied separately.

Absence (2005/06 onwards)

Prior to 2005/06, the main source of absence data was the annual Absence in Schools Survey (DfE 2011). However, since 2005/06 absence data has been returned within the school census as part of the termly return. The coverage of different settings/types of maintained schools, the age groups of children, as well as the breadth of indicators included in the

[2] Including: maintained nursery, primary, secondary, middle-deemed primary, middle-deemed secondary, local authority maintained special and non-maintained special schools, academies including free schools, studio schools and university technical colleges and city technology colleges. Independent (private) schools, and those who are home-schooled are not included.

absence data have increased over time. There are gaps in these data, documented by the DfE - for example, the pupil referral unit 2012/13 return does not include absence data (DfE 2019a).

Exclusions (2001/02 to 2005/06 and 2006/07 onwards)

Exclusions data have been collected as part of the school census from 2001 onwards. During the years 2001/02 to 2004/05 only a limited amount of information was recorded: whether the pupil had received a permanent exclusion and, if yes, the start date of this exclusion. From 2005/06, the data includes information on both fixed and permanent exclusions, including the reason for the exclusion and the start and end dates (DfE 2017).

Pupil Referral Unit (PRU) Census (2009/10 – 2012/13)

The PRU Census was an annual statutory return for all PRUs and alternative provision academies, including alternative provision free schools, in England. Initially, a more limited return was required from PRUs, including school and individual-level demographic markers and provision, before a wider number of indicators were required termly as part of the school census return.

Early Years Census (EYC) (2007/08 onwards)

The individual level EYC reflects the variation in provision of the funded early years sector. It has more detail on the type and scope of provision, the qualification level of staff, the funding sources, staff and child numbers, as well as individual characteristics of the children.

2.1.1.2 Attainment Data

In state schools in England, key attainment data are statutorily returned and stored centrally throughout a student's learning journey. The attainment data contain a mixture of teacher assessments and test results. Regulated qualifications sat and awarded in the last years of compulsory schooling (Key Stage 4, Key Stage 5 and beyond) are managed by external awarding bodies. After age 14, qualifications can be taken by students who are not in state-maintained provision - for example those who are home-schooled or privately educated; therefore, the data from awarding bodies provides some detail, including personal identifiers, on students who may not be present in the school census or ILR returns.

Early Years Foundation Stage Profile (EYFSP) (2002/03 onwards)

The EYFSP is a teacher assessment of the individual child's achievements at the end of the early years provision (age 5; end of reception year). This is submitted towards the end of the summer term. The data are richer from 2005/06 and include all providers when the return became statutory from 2008. Before this time, the return included fewer characteristics. Data are based on a 10% sample until September 2006.

Phonics (2011/12 – 2017/18)

The phonics screening check assessment, completed by the teacher, is a formal assessment of a child's reading ability at the end of year 1, at around age 6. It is an annual statutory data

collection, submitted by the local authority each June, for all state-maintained schools, academies, free schools and special schools.

Key Stage 1 (KS1, 1997/98 onwards)

KS1 covers years 1 and 2, when the child is aged 5-7 years. KS1 assessments are a statutory annual data collection at the end of year 2. Local authorities submit KS1 assessments for all state-funded schools, academies and free schools. The data include teacher (summative) assessments for all children, with the addition of points scores for formal assessments (commonly called SATs) for those children working at around the expected level. The information collected and recorded has changed over time.

Key Stage 2 (KS2) (1995/96 onwards)

KS2 includes years 3 to 6, when the child is aged 7 to 11 years. KS2 assessments are a statutory annual data collection at the end of year 6, before entering secondary school. Local authorities and individual schools submit KS2 summative assessments for every pupil in all state-funded schools, academies and free schools. Formal assessments are delivered by the school following statutory guidance and submitted separately. The information collected has changed over time, but with assessments in English and mathematics for all years.

Key Stage 3 (KS3) (1997/98 – 2012/2013)

Teacher assessments at the end of KS3, usually when the child is in year 9 (aged 13 or 14 years), are made in the core subjects of English, mathematics and science and also in non-core subjects, such as geography, art and music. The central collection of results for non-core subjects and statutory tests both ceased in 2007/08, when formal testing for this Key Stage ended. Statutory returns for this age group ceased in academic year 2012/13.

Key Stage 4 (KS4) (2001/02 onwards)

At the end of KS4, when the child is ~16 years, most students complete some form of regulated qualification within their learning establishment. Most students now take General Certificates of Secondary Education (GCSEs) in core and non-core subjects. Students who do not take regulated qualifications[3] are not present in these data.

[3] For 2021, current qualifications are: Entry level award, entry level certificate (ELC), entry level diploma, entry level English for speakers of other languages (ESOL), entry level essential skills, entry level functional skills Skills for Life, GCSE - grades 3, 2, 1 or grades D, E, F, G, level 1 award, level 1 certificate, level 1 diploma, level 1 ESOL, level 1 essential skills, level 1 functional skills, level 1 national vocational qualification (NVQ), music grades 1, 2 and 3, CSE - grade 1, GCSE - grades 9, 8, 7, 6, 5, 4 or grades A*, A, B, C, intermediate apprenticeship, level 2 award, level 2 certificate, level 2 diploma, level 2 ESOL, level 2 essential skills, level 2 functional skills, level 2 national certificate, level 2 national diploma, level 2 NVQ music grades 4 and 5, O level - grade A, B or C, access to higher education diploma, advanced apprenticeship, applied general, AS level, international Baccalaureate diploma level 3 award, level 3 certificate, level 3 diploma, level 3 ESOL, level 3 national certificate, level 3 national diploma, level 3 NVQ, music grades 6, 7 and 8, tech level, certificate of higher education (CertHE) higher apprenticeship, higher national certificate (HNC), level 4 award, level 4 certificate, level 4 diploma, level 4 NVQ, diploma of higher education (DipHE), foundation degree, higher national diploma (HND), level 5 award, level 5 certificate, level 5 diploma, level 5 NVQ, degree apprenticeship, degree with honours - for example bachelor of the arts (BA) hons, bachelor of science (BSc) hons, graduate certificate, graduate diploma level 6 award, level 6 certificate, level 6 diploma level 6 NVQ, ordinary degree without honours, integrated master's degree, for example master of engineering (MEng), level 7 award, level 7 certificate, level 7 diploma, level 7 NVQ, master's degree, for example master of arts (MA), master of science (MSc), postgraduate certificate in education (PGCE), postgraduate diploma, doctorate, for example doctor of philosophy (PhD or DPhil), level 8 award, level 8 certificate, level 8 diploma.

Key Stage 5 (KS5) (2001/02 onwards)

KS5 is the last phase of compulsory schooling in England, ending when the child is 17 or 18. (before 2013, the compulsory school leaving age was at the end of KS4, aged 16). Data on qualifications are returned by the awarding bodies, learning providers and local authorities as part of the AAT. A range of overall progress (value-added) measures are available only for later years, and a flag indicates whether the same methods of calculating these can be applied to students leaving school before this.

2.1.1.3 Social Care Data

Children in Need (CIN) (2008/09 onwards)

CiN captures child-level information on children referred to and assessed by children's social care services (Emmott et al. 2019). In the first year, these data covered all children assessed between 1 October 2008 to 31 March 2009. From 2010 it widened to any child referred to children's social care services within the year, and any cases open on 1 April 2009. There are gaps in these data. For example, 151 out of 152 local authorities submitted the CiN census for 2009/10, of which a number only submitted the aggregate data due to quality concerns at the individual level. There are further quality concerns with the 2009/10 data; these are reported within the OSR24 statistical release documentation (DfE, 2010).

Children Looked After (CLA) (2005/06 onwards)

The yearly CLA return is a local authority return on all children in local authority-maintained care and recent care leavers within the financial year 1 April to 31 March. Although data collection started in 1992, between 1998 and 2003 it was restricted to a one-third sample. (McGrath-Lone et al. 2016). The linked dataset contains data from 2005/2006 onwards and only includes the CLA data that are incorporated into the NPD, which only covers the most recent episode of care in a given year.

Note that, for our project, we separately requested the full episode CLA data (SSSA903 return; DfE 2019b), which includes all episodes and contains information on the start and end date of each episode, the category of need, the type of placement and the reason for a placement change, if applicable.

2.1.1.4 Further education and destination data

Individualised Learner Record (ILR)

The Individualised Learner Record, or ILR, is a monthly return of data from further education and work-based education learning providers, which includes individual records of students aged 14 and above. A limited set of individual-level data is incorporated into the Young Person's Matched Administrative Dataset (available in the current share of the MoJ-DfE dataset, see below), and others, depending on the type of institution and the year of the entry.

Post-16 Learning Aims (PLAMS) (2007/08 onwards)

This dataset contains detailed information about learning aims and achievements, including entry and exit markers. PLAMS data are returned within the school census by schools who have sixth form provision, within the ILR, and by awarding bodies.

National Client Caseload Information System (NCCIS) (2010/11 onwards)

Since September 2013, all 16-year-olds in England have been legally required to participate in education or training until the age of 18. The NCCIS is an annual local authority return from the local version of the administrative database that was put in place to record young people's engagement in education and training, to identify those who are not participating, and to plan services.

Young Person's Matched Administrative Dataset (formerly Level 2 and 3 at 19 indicators) (2017/18)

This is a matched dataset containing information on the attainment of young people who were aged 19 in the academic year 2017/18. Data collected on individuals registered on YPMAD includes pupil, registration and establishment details and information about attainment and apprenticeships from the pupil level annual school census, ILR, and post 16 qualifications data collections (AAT).

Independent Specialist Providers (2009 – 2013)

This student-level annual return is from state-funded independent specialist providers of education and training to learners with learning difficulties and/or disabilities who are over 16, but under 25, and who are subject to a learning difficulty assessment. Data are now returned through the ILR or the school census.

2.2 Ministry of Justice data

The MoJ process record-level administrative criminal justice data in England and Wales, including cautions, convictions, sentencing, defendant outcomes, custodial (prison) details and discharge and reoffending data. Common to all administrative data systems, there have been changes over time in the way that these data are collected. The MoJ reports on revisions, data sources, quality and dissemination, and methodological developments (MoJ, 2019 and 2020).

2.2.1 Police National Computer (PNC) (2000 – 2017)

These data originate from the recording system used by law enforcement agencies throughout the UK for recordable crime at the time of the event and, as such, the original source includes the personal identifiers used for linkage, as well as subject-reported or recorder-reported ethnicity. Although the live PNC is 'weeded' for various reasons, the MoJ PNC, which is primarily used for research purposes, is weeded for deduplication. Therefore, a conviction at 15 for a current 19-year old would remain in the data.

The data included in the MoJ-DfE linked dataset include information on all offenders who were convicted or cautioned for a recordable offence in England. The MoJ PNC data includes the information necessary to establish individuals' criminal histories, but does not include arrest data. The data extracted goes up to the end of 2017, and includes the offence, the disposal type (or outcome, such as community service), and the sentence received (if applicable).

2.2.2 Home Office Court Appearance Statistics (HOCAS) (2009 – 2017)

HOCAS data originates from the magistrates' court case management system ('Libra') reports, which are hosted on Her Majesty's Courts and Tribunals Service (HMCTS) Performance Database ('OPT'). The reports cover all cases, criminal or otherwise, dealt with in the magistrates' courts. The data provided by the courts are checked and verified at case level by court staff before being submitted on OPT. HMCTS staff perform further checks, including the investigation of any anomalies, missing data returns and any unexpected changes in the data (MoJ, 2020)

This dataset includes defendant outcomes. Personal identifiers may be passed from the PNC to the courts if the offence was processed by a law enforcement agency. For some offences - for example motoring and credit or license default - details may be passed from the relevant agency directly to the court, where they would be entered into the HOCAS data.

2.2.3 Case management system for crown court cases (CREST) (2008 – 2017)

Details shared in the extract of this dataset include date and type of offence, the number of offenders within the case, date of the hearing, and the recorded outcome.

2.2.4 Prison data

From 2008 to 2015 the data from individual prison establishments were recorded on the Prison-NOMIS system which feeds through to a central computer database (Inmate Information System (IIS)), from which data extracts are used to produce the various prison datasets listed below. The same processes were used to extract data on all offenders between 2005 to 2015. From 1st January 2015, the data extracts used to produce statistics on prison receptions transitioned to a new extract, which is taken from the Prison-NOMIS system directly and without needing to be processed by the IIS. The same transition came into effect on 30th June 2015 for the data extracts used to produce statistics on the prison population. The new extract has more accurate sentence length information and richer detail about offences committed. In October 2015, this transition was implemented for release data. The MoJ-DfE linked data includes:

- Prison population old dataset (January 2007 – June 2015)
- Prison new population dataset (July 2015 - Dec 2017)
- Prison discharge old dataset (January 2008 – December 2014)
- Prison discharges new dataset (2015-2017)
- Prison receptions old dataset (January 2008 - Sep 2015)
- Prison new receptions dataset (2015-2017)

3 Data linkage

The UK has no population register with mandatory coverage of all residents, and thus no corresponding resident unique ID number to facilitate the linkage of records from across government departments. There are also legal barriers to the systematic linkage of departmental identifiers to create persistent and sustained 'bridging IDs' spanning departments. Due to these factors, an exercise to link education and justice data requires novel data linkage process involving considerable reconciliation of information (in a pre-linkage cleaning process) and subsequent assessment of linkages between the two populations. This section describes the processing of the data through the cleaning and linkage process. The linkage process was determined and implemented by the MoJ and DfE.

Target population (denominator)

In the linkage exercise, the DfE included data on all individuals with identifiers in the school census data, attainment data, the ILR and the Higher Education Statistics Authority (HESA) data (see Section 3.3). The aim was to include individuals educated for at least one period between the ages of 2-18 in England. This would include all individuals who were educated in a state-maintained setting or independent specialist provider at some point during their schooling, individuals educated at independent schools or home-schooled who sat regulated exams, and individuals entering the higher education system and resident in England (we have assumed that HESA filtered their data to only include identifiers for students with home addresses in England, but this has not been verified). Some individuals in this population would have very little or no data in the NPD (for example, those whose identifiers came from the HESA data, those who attended an independent school for the majority of their schooling, and those who moved away from England early on in their schooling).

3.1 Internal linkage: DfE

Within the NPD, data are linked internally using a unique reference number (Pupil Matching Reference, PMR), generated as part of the matching process and based on the Unique Pupil Number (UPN), allocated on entering the system - for example when a child enters a state-maintained school, or from their Unique Learner Number (ULN) if entering education (in England) for the first time from age 14 onward. To note, children who are educated in an independent school and who to sit GCSEs will enter the system at this point. Schools and other providers share data records (including the UPN) when the child moves to their next destination - for example when a child moves from nursery to infant school, changes school, or sits a formal qualification. There are a series of documented validation checks performed by the provider or local authority submitting the data to reduce the possibility of a child or student having multiple UPNs.

3.2 Internal linkage: MoJ

The justice administrative systems, covering the whole of the UK, were not initially designed for direct linkage. Identifiers in the justice data are manually entered by either court or law enforcement staff. Records from across the justice system have been linked retrospectively by the creation of a linked individual ID. Historically, one person may have been allocated different unique IDs due to the way the IDs were allocated. These missed matches occur in the PNC data, and are updated throughout the year when new information becomes available and, as such, these are 'living' data.

As part of the Data Improvement Project, the MoJ developed the current internal linking methodology (MoJ 2011) and has been performing regular data linkage projects across the justice sources to improve linkage quality. In 2019, all records up to the end of 2017 were included from all core criminal justice data sources (excluding TAR, magistrates final appearances data) and were re-linked using deterministic methods[4], quality assessed and assigned new unique linked IDs. The linked IDs were used to prepare the extract from the Police National Computer (PNC) and magistrates' court records from HOCAS that were shared for linkage. Selecting these two sources offered the most reliably recorded and widest coverage of offending history.

To decrease the possibility of missed matches due to missing individual identifiers in the justice data, a number of cases included multiple birthdates and/or alias names. The proportion of offenders, identified by their unique linked ID, with multiple birthdates in either the PNC or HOCAS data was less than or equal to 4%.

3.3 Identifiers used for linkage

The identifying characteristics used for linkage were first name, last name, date of birth, gender and postcode and originated from six sources – four DfE sources (the school census, attainment, ILR and the HESA dataset[5]) and two MoJ sources (the PNC and HOCAS data). The relative order of importance for these sources was based on typical data quality and was determined by the DfE and the MoJ, in descending order, as: school census, attainment, ILR and HESA for the DfE data; and PNC and HOCAS for the MoJ data. To note, personal identifiers are not necessarily static across time; it is therefore possible for an individual to have different identifiers at different timepoints.

[4] Outwith the MoJ-DfE project, the MoJ also use the Fellegi-Sunter/Expectation Maximisation framework, a blocking routine and a probabilistic matching set of algorithms; they built on previous work in this area by creating open-source software, which is now documented with updates (**https://github.com/moj-analytical-services/splink**).

[5] HESA are the controller of individual level information about publicly funded Higher Education in the UK. As such, they hold personal identifiers for each student, including their Unique Pupil Number from the DfE. If any field (obtained from ILR or school census) is blank, HESA may import these data directly from a potential student's Universities and Colleges Admissions Service (UCAS) application to university. This set of personal identifiers was used for matching, but no HESA data were included in the linked dataset.

3.4 Cleaning identifiers prior to linkage

In February 2020, the MoJ pre-prepared an individual-level dataset containing the personal identifiers and the MoJ linking ID, and shared this with the DfE. The MoJ linking ID was the same for all records relating to an individual. The MoJ shared 2,631,842 individual cases (including 85,233 cases with multiple birthdates).

On receipt of the MoJ data, the DfE linkage team performed manual cleaning routines on first and last names in both the MoJ and DfE datasets. These routines are standard practice based on their previous linkage projects. The MoJ name fields were standardised to match the DfE data sources. Cleaning included: left-right trimming to remove spaces, hyphens, and accents, checks for concatenation (for example first and middle name joined together in one field), correct ordering (for example, date of birth not transposed, where entries are not plausible), double entry names (Jon/Jonathan etc.), first and last names in the correct order and in the correct box. No cases were deleted before linkage. This step is undocumented; as such, we do not know, for example, whether dates of birth were reconciled, if a lexicon was used for names, or whether the addresses were verified with a gold standard dataset such as AddressBasePlus.

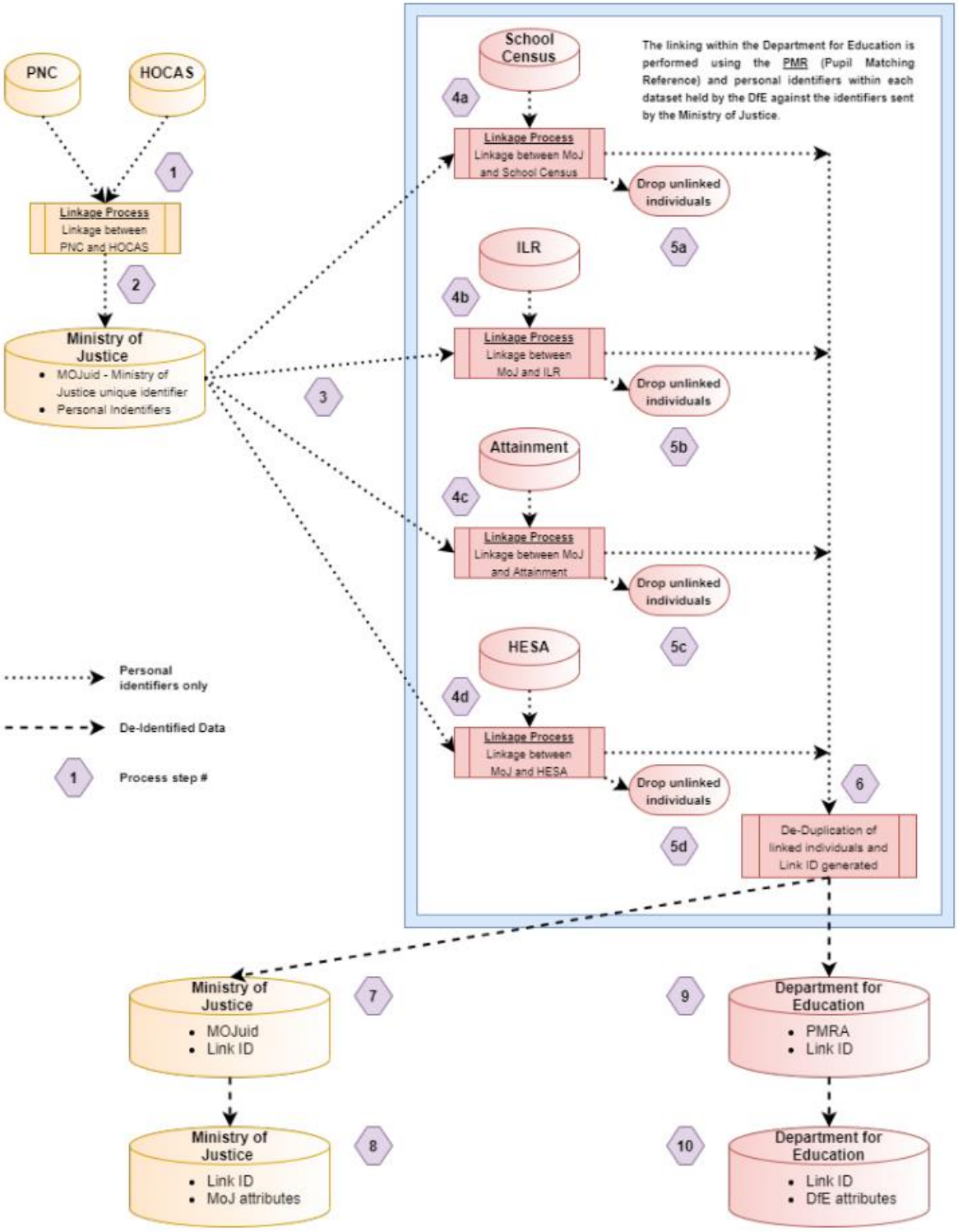
3.5 The linkage process

The data flow steps, illustrated visually in Figure 3.1, were:

1. The MoJ linked the PNC and HOCAS datasets by adding any HOCAS individuals to the PNC set.
2. The MoJ created a unique identifier for the study (MoJUID) against all their personal identifiers.
3. The MoJ identifier dataset was passed to the DfE.
4. In steps 4a) to 4d) the DfE linked the MoJ dataset against the identifiers held in each of their four main datasets.
5. Unlinked individuals were dropped from each linkage sub-stage. All were matched to 4a) and then all to 4b) and then all to 4c) and then all to 4d).
6. The four datasets containing MoJ and DfE linked identifiers were then de-duplicated to create a master Link ID between the MoJUID and the PMR.
7. The MoJUID and Link ID were returned to the MoJ without any personal identifiers.
8. The MoJ attribute (activity) dataset was constructed as an anonymous dataset using the Link ID.
9. The PMR and Link ID were returned to the DfE without any personal identifiers.
10. The DfE attribute dataset was constructed as an anonymous dataset using the Link ID.

Further details on parts of this process are given in Section 3.5.1.

Figure 3.1: Data flow diagram showing the slow of personal identifiers during the linkage process



3.5.1 Linkage algorithm

This refers to process steps 4a to 4d listed above and shown in Figure 3.1. The linkage algorithm used was deterministic. This is a rule-based method where a set of personal identifiers are compared across the data and, if they all agree, the records are linked. The DfE employed a manual iterative deterministic approach to linkage, using increasingly less restrictive versions of the individual identifiers, with 103 steps divided into 7 descending matching priorities. This process was performed by the linkage team using SAS 15.1. These steps and priorities are called matching rules. Due to personal identifiers not being wholly unique to each person, or even constant over time, each of the steps is designed to account for some recording error or mismatch (for example, where the first names and last names, or the month and year of birth, have been entered in reverse order in one record and not in another). Each of these steps was performed on each of the four DfE sources separately. The full set of matching rules and priorities are given in Appendix C, but are summarised in Table 3.1.

Table 3.1: Description of the hierarchy of matching rules applied

Match priority	Description
1	Exact names (including middle names), date of birth (DOB), postcode, gender
2	Relaxing name fields and spelling or gender or DOB separately
3	Relaxing match priority 2 by adding one additional rule change at a time
4	Relaxing match priority 3 by using less restrictive combinations of two rules
5	Relaxing postcode to three characters without gender or with a fuzzy name
6	Two-character postcode for where there is a 1:1 match on first name, last name and DOB
7	1:1 match on first name, last name and DOB

The linkage algorithm was as follows:

Step 1: Linkage by the DfE

For each matching rule, the DfE ran matches for each MoJUID against each of the data sources. All match possibilities (1 MoJUID:1 PMR; 1 MoJUID:many PMRs; many MoJUIDs:1 PMR; many MoJUIDs:many PMRs) were permitted. If two records matched on the first pass, they were not excluded for the second, as all matches were considered.

Step 2: Clerical review by the DfE

Fuzzy sets, where partial use of identifying characteristics from match priority 2 and below (Table 3.1) were reviewed by DfE data preparation staff not performing the linkage. Individual matches were accepted or rejected using rules used in previous external linkage projects as well as professional judgement. For example, linkages where gender was not recorded, where date of birth digits were transposed, or where common names introduced the possibility of error, would require professional judgement based on the information within the data.

The clerical review process was not described – for example in a protocol or set of standard operating procedures – and thus could not be assessed. Results of the clerical review were not recorded or quantified – and, as such, were not available for review.

Step 3: Match clean and reporting by the MoJ

Prior to cleaning, the MoJ reported on matched and unmatched cases by key variables within the justice data. These are reported in Section 4.

The DfE shared the sets of matched PMRs and MoJUIDs with the MoJ, dropping the identifiers, but including a set of flags to show the source datasets and the matching priority for each matched individual. Where one MoJUID matched with one PMR, the record was retained.

For the remaining cases, the MoJ used the following selection criteria:

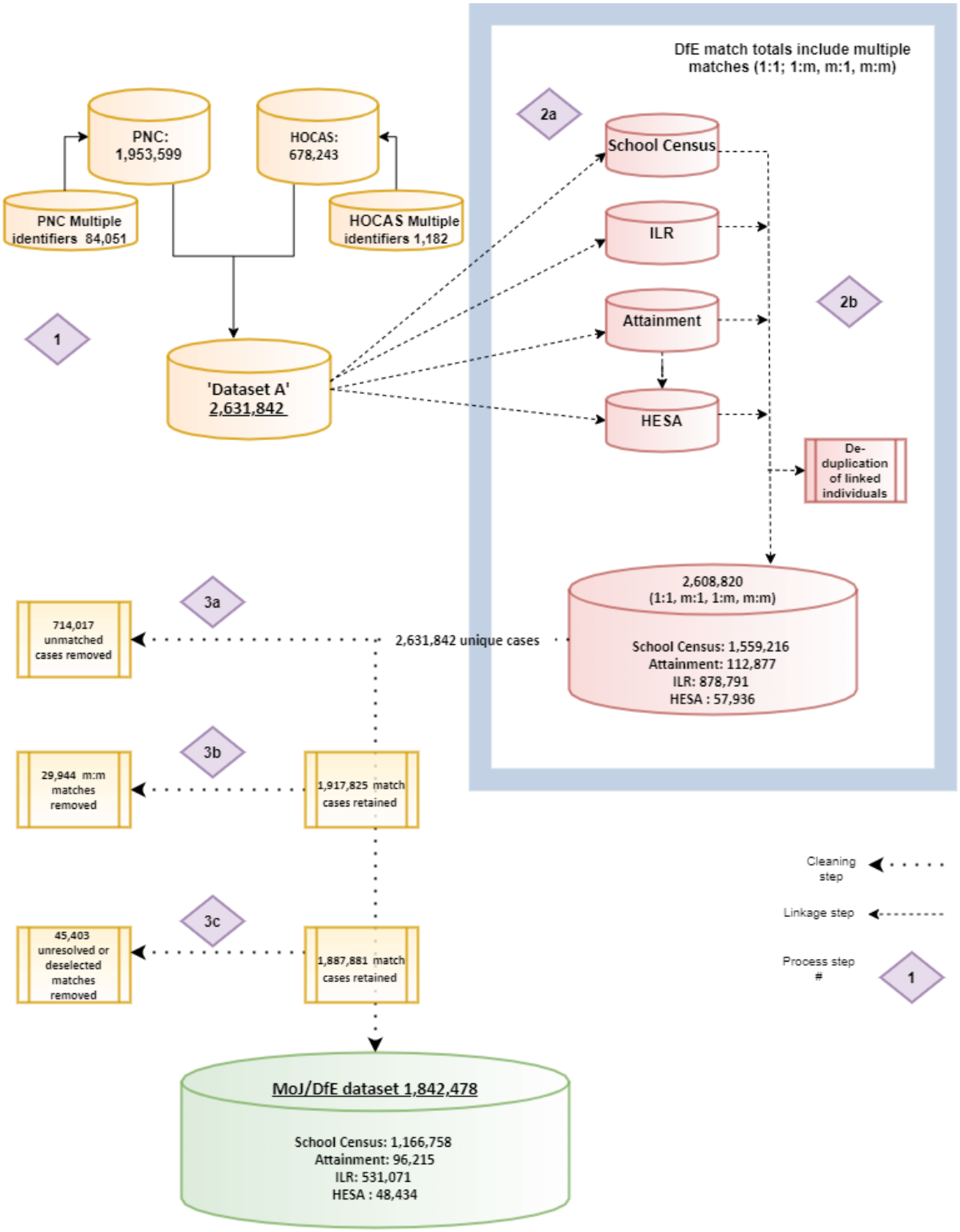
- For single MoJUIDs matched to multiple PMRs, the MoJ selected the most reliably matched PMR as follows:
    - Using the strictest matching rule
    - Where multiple PMRs shared the same strictest matching rule, prioritising the DfE datasource: 1. School census, 2. Attainment, 3. ILR and 4. HESA
    - Less strict pairs of records were dropped
    - Multiple PMRs which shared the same matching rule and DfE data source were dropped
- For single PMRs matched to multiple MoJUIDs, the MoJ selected the most reliably matched MoJUID as follows:
    - Using the strictest matching rule
    - Prioritising the MoJ data source: 1. PNC and 2.HOCAS
    - Prioritising the DfE data source (as above)

Many-to-many matches still occurred, where IDs from one source matched to IDs from the other source, and the matching rule and data source were the same. These were quantified, but not fully reported in terms of source, and were dropped from the final data tables.

Step 4: Post-match reporting by the MoJ

The MoJ reported on the aggregate proportions of retained and removed cases based on justice-recorded characteristics (gender, ethnicity, year of birth and offence and others).

Figure 3:2: Flow diagram detailing the number of individuals with an offending history who were linked in the data



3.6 Linkage results

From the MoJ, identifiers associated with 2,631,842 MoJUIDs were submitted for matching (Step 1). We do not know how many individuals were in the DfE dataset of identifiers or whether these were restricted on dates of birth because this was not recorded.

The DfE linkage process (Figure 3.2, Step 2) and subsequent cleaning based on the source of the identifier by MoJ (Step 3) resulted in a final 1:1 match between the combined MoJ dataset and at least one of the four DfE datasets for 70% of individual cases in the MoJ dataset, meaning that 1,842,478 unique individuals with at least one record in the criminal justice data were matched against a single unique pupil record. Of the original 2,631,842 individuals submitted by the MoJ for matching (Step 1), 714,017 (27%, Step 3a) of unique individuals within the MoJ dataset were not matched to any record in any of the education datasets. These unmatched cases were not included in the dataset.

The MoJ data team then removed 29,944 (1.1%) cases where many IDs had linked to many alternative IDs (Step 3b). From the remaining 45,403 (Step 3c), 40,324 cases (1.5% of the original number of MoJ cases) were removed, as they were not the best match available, based on the source files used for matching. The remaining 5,079 of removed individuals were matches where the source files were equal, but the matches were different. Therefore, the linkage process identified a total of 1,842,478 unique individuals with a recorded criminal justice outcome with their corresponding education records.

These results are summarised in Figure 3.2. Among the matched individuals, 1,182,841 (64% of the final matched individuals, or 45% of the original 2,631,842) were matched on the strictest match priority (Table 3.2).

Table 3.2: Match rate for the hierarchy of matching rules

Match priority	Description	% Matched
1	Exact names (including middle names), date of birth (DOB), postcode, gender	64%
2	Relaxing name fields and spelling or gender or DOB separately	15%
3	Relaxing match priority 2 by adding one additional rule change at a time	6%
4	Relaxing match priority 3 by using less restrictive combinations of two rules	2%
5	Relaxing postcode to three characters without gender or with a fuzzy name	0.03%
6	Two-character postcode for where there is a 1:1 match on first name, last name and DOB	2%
7	1:1 match on first name, last name and DOB	11%

4 Linkage comparison

All tables and figures in this section (Sections 4: 4.1 and 4.2) use data from the MoJ's bias testing report (MoJ, 2021b) The analysis was carried out by the MoJ.

4.1 Comparison of matched and unmatched cases

The MoJ produced the following analyses on the PNC source comparing characteristics of matched and unmatched individuals and records (MoJ, 2021b) This was carried out on the dataset after the DfE's clerical review (after Step 2), and prior to the MoJ's resolution of the one-to-many, many-to-one and many-to-many matches (i.e. not on the final set of matched and unmatched individuals) but did not include the individuals from the MoJ data sources who were entered with multiple dates of birth (ibid.).

4.1.1 Gender

The gender distribution was similar in the matched and unmatched PNC data (Table 4.1).

Table 4.1: Gender of matched and unmatched cases

Gender	PNC	
	Matched	**Unmatched**
Male	1,132,221 (73%)	301,398 (75%)
Female	413,670 (27%)	96,968 (24%)
Unknown	6,522 (<1%)	2,691 (<1%)

4.1.2 Age (year of birth and age at offence)

Unmatched cases were older, on average, than matched cases. The MoJ believes that the increased likelihood of matching for younger offenders is likely to be a result of using postcode in the matching algorithm, with the younger offending population more likely to report living at the same address as the one recorded by the learning provider.

Figure 4.1: Year of birth in matched and unmatched cases

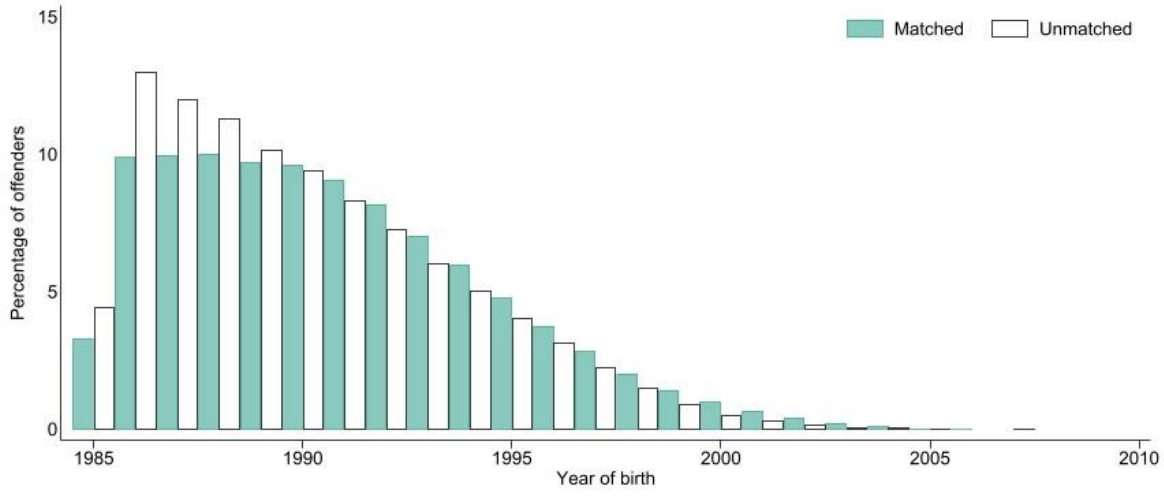
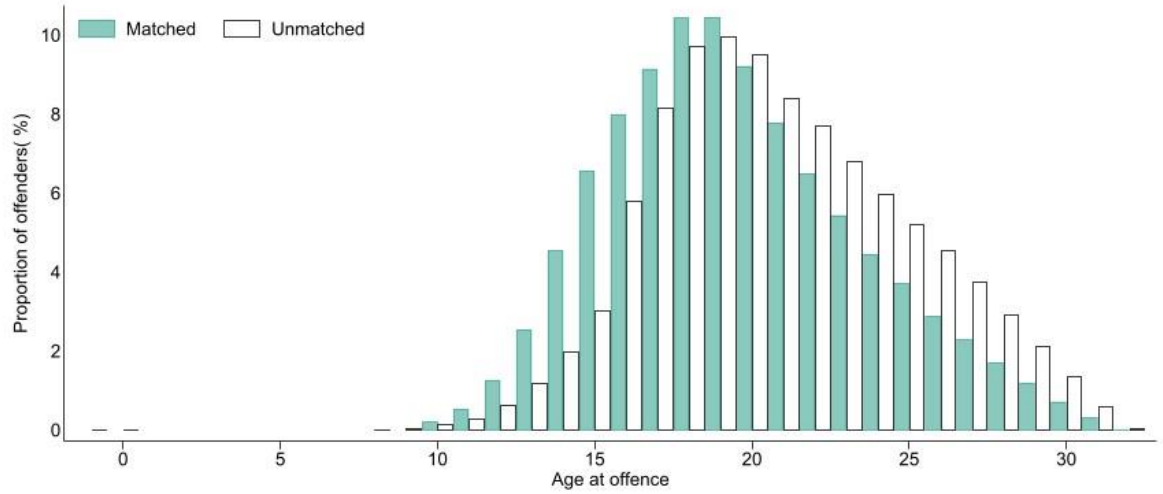


Figure 4.2: Year of birth in matched and unmatched cases



4.1.3 Ethnicity

Ethnicity (officer identified) is only recorded in the PNC data. Proportions of White – North European, Black, and Asian individuals were higher in the matched data whereas those recorded as White – South European, SE Asian, Middle Eastern, and unknown were higher in the unmatched data (Table 4.2).

Table 4.2: Ethnicity of matched and unmatched cases

Ethnicity	Matched	Unmatched
White – North European	1,238,985 (80%)	298,895 (75%)
White – South European	22,144 (1%)	19,396 (5%)
Black	126,150 (8%)	12,334 (3%)
Asian	84,262 (5%)	15,577 (4%)
Chinese, Japanese or SE Asian	4,699 (<1%)	4,723 (1%)
Middle Eastern	8,500 (<1%)	4,668 (1%)
Unknown	66,423 (<1%)	45,162 (11%)

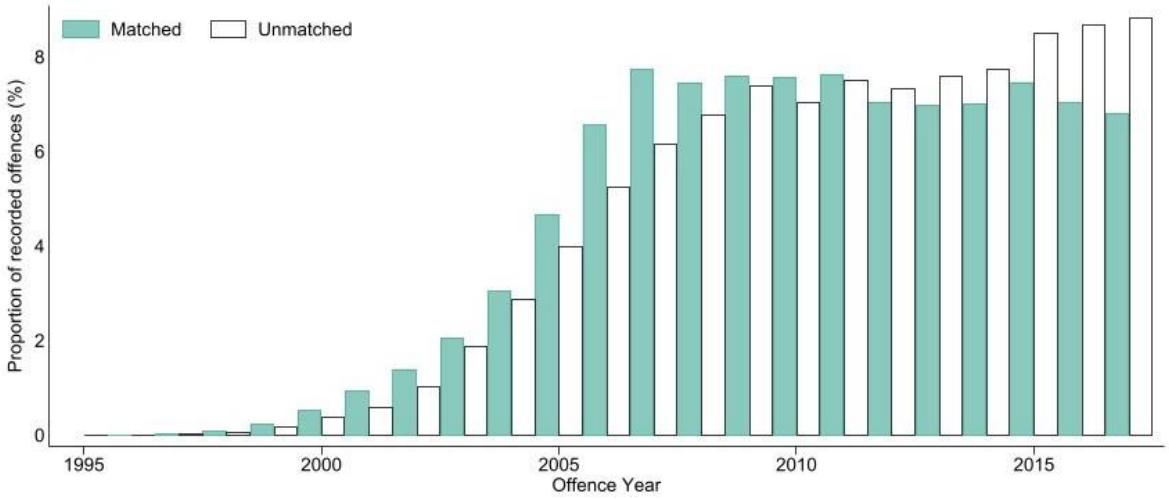
4.1.4 Offence year

Offences in the unmatched data were, on average, more recent than in the matched data. The results for the PNC are given in Figure 4.3.

4.2 Comparison of retained and removed cases

As described in Step 4 of the linkage algorithm (Section 3.5.1), m:m matches were removed, the preferred 1:1 record was selected from the m:1 and 1:m matches using the hierarchy of data sources, and the alternatives were removed. Where the 1:m and m:1 matches could not be reconciled by the hierarchy of source data, the matches were removed. Further comparisons were carried out after post-match cleaning, comparing retained and removed cases and records. The retained cases below include those who had a 1:1 match at Step 1.

Figure 4:3: Year of offence in matched and unmatched cases



4.2.1 Gender

Males and individuals with gender recorded as 'unknown' made up a higher proportion of the group removed.

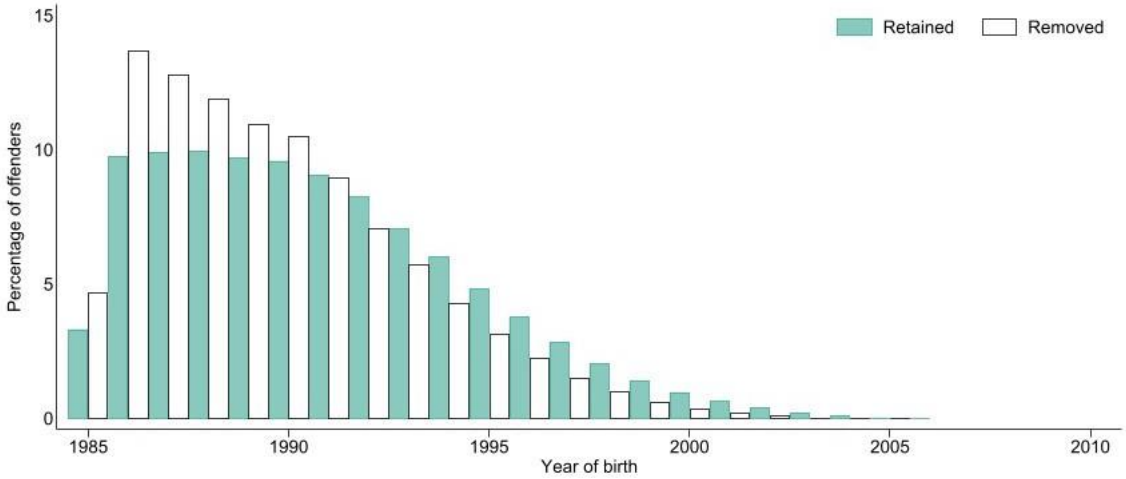
Table 4.3: Gender of retained and removed cases

Gender	PNC	
	Retained	**Removed**
Male	1,099,056 (73%)	32,853 (81%)
Female	406,657 (27%)	6,912 (17%)
Unknown	5,893 (<1%)	629 (2%)

4.2.2 Year of birth

Removed cases were older, on average, than retained cases (Figures 4.4 and 4.5).

Figure 4:4: Year of birth of retained and removed cases



4.2.3 Ethnicity

Black and Asian individuals made up a higher proportion of the records removed by the cleaning than those retained (Table 4.4).

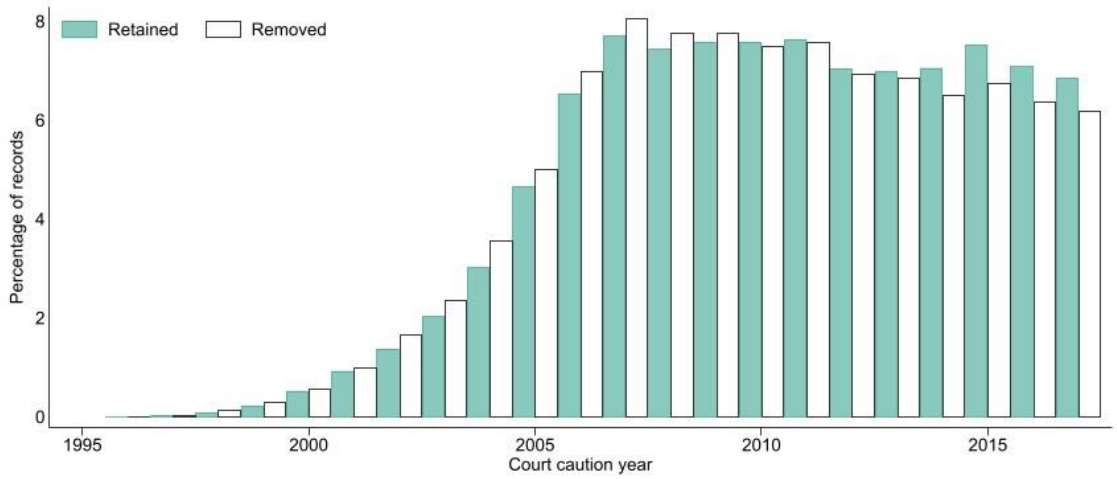
Table 4.4: Ethnicity of retained and removed cases

Ethnicity	Retained	Removed
White – North European	1,212,826 (81%)	26,159 (65%)
White – South European	21,565 (1%)	579 (1%)
Black	119,1548 (8%)	6,602 (17%)
Asian	80,214 (5%)	4,048 (10%)
Chinese, Japanese or SE Asian	4,572 (<1%)	127 (<1%)
Middle Eastern	8,107 (<1%)	393 (1%)
Unknown	64,245 (4%)	2,178 (5%)

4.2.4 Year of offence

The retained records were, on average, slightly more recent than removed records (Figure 4.5).

Figure 4:5: Year of offence in the retained and removed cases



PART TWO: Data quality

5 Part two introduction

In this part we report on quality checks carried out on the data. We were granted access to a subset of datasets and variables in the linked data. Specifically, from the MoJ we requested data from the PNC, HOCAS and CREST data. However, CREST data were not accessible at the time of data extraction due to Covid-19 restrictions so were not included in our extract.

From the DfE data we requested data from the following datasets: pupil-level school census, PRU census, AP census, early years census, early years foundation stage profile, Key Stage 1 attainment, Key Stage 2 attainment, Key Stage 3 attainment, Key Stage 4 attainment, Key Stage 5 attainment, absence, exclusions, NCCIS, CiN and CLA. We also requested access to the linked CLA return (SSDA903), which contains full episode CLA data, and school-level census and attainment data.

We could not assess all variables in all datasets, so predominantly evaluated key variables from the main datasets likely to be used by researchers for the period from reception (age 4/5) up to the end of Key Stage 4 (age 16). Thus, we did not assess any Key Stage 5 data or any data from the NCCIS dataset. The main exception to this is in part of the report where we evaluate completeness in some variables from the pupil-level census, in which we take all school years from reception onwards. We do not report on any individuals appearing only in the HOCAS data (i.e. in the HOCAS dataset but not in the PNC), except in the first section on match quality.

This part of the report has five sections. Section 6 summarises the data on match quality, Section 7 gives a summary of the numbers in the linked dataset, and Section 8 gives an overview of the data available for which school years (and birth cohorts). In Section 9 we report on completeness in key variables and in Section 10 on consistency and uniqueness.

6 Match quality

Altogether, in the linked data there were 1,842,478 individuals with an MoJ unique identifier (MoJUID) and matched Pupil Matching Reference (PMR). Tables 6.1 and 6.2 give the MoJ and DfE sources for the matching identifiers, together with the matching rule applied. As would be expected given the selected hierarchy of source data, matching rule 1 (exact names, including middle names, date of birth (DOB), postcode and gender) was more likely to have been applied when the identifiers came from the PNC data (69.4% of matches) compared to when they came from the HOCAS data (40.4% of matches). Matching rules 6 or 7 were more likely to have been applied when the (MoJ) identifiers came from the HOCAS data (4.4% for rule 6 and 13.6% for rule 7 for HOCAS compared to 2.0% and 10.1%, respectively, for the PNC).

Similarly, matching rule 1 was more likely to have been applied when the (DfE) identifiers came from the school census (71.5% of matches) or the ILR (64.2% of matches) compared to when they came from attainment data (0.8% of matches) or from the HESA data (14.7% of matches). Where the identifiers came from attainment data, 98.9% of matches used matching rule 7 (1:1 match on first name, last name and DOB).

Table 6.1: Source of MoJ identifiers and matching rule applied for all matched individuals from the PNC or HOCAS data

Matching rule	**MoJ data source for identifiers**		
	PNC	**HOCAS**	**All**
1	1,049,367 (69.4%)	133,474 (40.4%)	1,182,841 (64.2%)
2a	15,051 (1.0%)	19,433 (5.9%)	34,484 (1.9%)
2b	75,228 (5.0%)	19,527 (5.9%)	94,755 (5.1%)
2c	87,584 (5.8%)	38,291 (11.6%)	125,875 (6.8%)
2d	9,969 (0.7%)	4,148 (1.3%)	14,117 (0.8%)
2e	391 (<0.1%)	339 (<0.1%)	730 (<0.1%)
3a	32,671 (2.2%)	13,374 (4.1%)	46,045 (2.5%)
3b	1,040 (<0.1%)	1,630 (0.5%)	2,670 (0.1%)
3c	5,617 (0.4%)	8,004 (2.4%)	13,621 (0.7%)
3d	257 (<0.1%)	775 (0.2%)	1,032 (<0.1%)
3e	8,146 (0.5%)	7,178 (2.2%)	15,324 (0.8%)
3f	1,112 (<0.1%)	764 (0.2%)	1,876 (0.1%)
3g	1,587 (0.1%)	1,285 (0.4%)	2,872 (0.2%)
3h – 3k	152 (<0.1%)	223 (<0.1%)	375 (<0.1%)
3l, 3o	19,983 (1.3%)	5,402 (1.6)	25,385 (1.3%)
4a	14,608 (1.0%)	6,210 (1.9%)	20,818 (1.1%)
4b	587 (<0.1%)	1,200 (0.4%)	1,787 (0.1%)
4c	4,571 (0.3%)	5,479 (1.7%)	10,050 (0.6%)
4d	565 (<0.1%)	527 (0.2%)	1,092 (<0.1%)
4e - 4g	20 (<0.1%)	63 (<0.1%)	83 (<0.1%)
5a	1,629 (0.1%)	3,209 (1.0%)	4,838 (0.3%)
5b	245 (<0.1%)	310 (<0.1%)	555 (<0.1%)
6a	29,438 (2.0%)	14,583 (4.4%)	44,021 (2.4%)
7a	152,288 (10.1%)	44,944 (13.6%)	197,232 (10.7%)
All	1,512,106	330,372	1,842,478

Table 6.2: Source of DfE identifiers and matching rule applied for all matched individuals from the PNC or HOCAS data

Matching rule	DfE data source for identifiers			
	School census	Attainment	ILR	HESA
1	833,655 (71.5%)	808 (0.8%)	341,245 (64.2%)	7,133 (14.7%)
2a	5,856 (0.5%)	2a-2e:	7,882 (1.5%)	20,734 (42.8%)
2b	57,790 (5.0%)	134 (0.1%)	36,364 (6.8%)	561 (1.2%)
2c	82,342 (7.1%)		40,290 (7.6%)	3,177 (6.6%)
2d	7,734 (0.7%)		6,226 (1.2%)	2d-2e:
2e	272 (<0.1%)		450 (<0.01%)	149 (0.3%)
3a	25,799 (2.2%)	3a-3o:	19,862 (3.7%)	380 (0.8%)
3b	652 (<0.1%)	83 (<0.1%)	893 (0.2%)	1,124 (2.3%)
3c	2,117 (0.2%)		4,453 (0.8%)	7,051 (14.6%)
3d	104 (<0.1%)		490 (<0.01%)	425 (0.9%)
3e	10,549 (0.9%)		4,476 (0.8%)	296 (0.6%)
3f	897 (<0.1%)		913 (0.2%)	53 (0.1%)
3g	1,726 (0.2%)		1,030 (0.2%)	110 (0.2%)
3h – 3k	90 (<0.1%)		249 (<0.01%)	35 (<0.1%)
3l, 3o	18,080 (1.5%)		6,963 (1.3%)	300 (0.6%)
4a	11,834 (1.0%)	4a-6a:	8,748 (1.7%)	233 (0.5%)
4b	339 (<0.1%)	20 (<0.1%)	607 (0.1%)	841 (1.7%)
4c	6,926 (0.6%)		2,966 (0.6%)	156 (0.3%)
4d – 4g	446 (<0.1%)		686 (0.1%)	39 (<0.1%)
5a	4,189 (0.4%)		159 (<0.01%)	490 (1.0%)
5b	555 (<0.1%)		0	0
6a	25,286 (2.2%)		18,185 (3.4%)	539 (1.1%)
7a	69,520 (6.0%)	94,720 (98.9%)	28,384 (5.3%)	4,608 (9.5%)
All	1,166,758	95,765	531,521	48,434

7 Numbers

Figure 7.1 shows the numbers in the dataset provided to us. Note that we were provided with NPD data for any individuals who had been allocated a pupil matching reference (i.e. all individuals in the NPD). As such, this included individuals born before 1985 as well as individuals born after 2007. The former group (those born before 1985) were not included in the data shared by the MoJ, so we have excluded these from further analyses. Similarly, those born after 2007 would not have a record in the PNC and are also excluded from any further analyses, except in the flowchart below.

The NPD datasets we accessed contained data on 23,371,459 individuals. The PNC data we accessed contained at least one record on 1,510,829 individuals. This included 57,764 individuals who were in the PNC but did not appear in the NPD datasets we accessed. For most (98%) of these individuals, the DfE data source (for identifiers) was the ILR or HESA data; given that only limited ILR (via the YPMAD) and no HESA data are included in the final dataset, this is not unexpected. For 99% of the 919 individuals whose source was either the school census (n=811) or attainment (n=108) data, the matching rule was 7a.

Thus, we started with a dataset that included 23,431,223 individuals (Figure 8.1). This included 1,282,246 individuals whose year and month of birth (according to the NPD) was before September 1985 (of these, 990 had a record in the PNC data – so, according to the MoJ data, would not have year and month of birth before September 1985) and 6,770,675 whose year and month of birth was after August 2007 (of these, 10 had a record in the PNC data).

Among the 990 who were excluded because their NPD year and month of birth was before September 1985, the DfE source for identifiers was the ILR for 817 (83%), the school census for 96 (10%), attainment data for 56 (6%) and HESA data for the remaining 21 (2%). Among the remaining 15,319,370 individuals, year and month of birth was missing for 289,247 individuals (3,877 of these had a record in the PNC and, of these, the ILR was the DfE source of data for identifiers for 3,583 (92%)). These 289,247 individuals with no information on year or month of birth in the NPD had no school census data for any school years. This left 15,029,291 individuals, of whom 1,446,188 (9.6%) had at least one record in the PNC.

Figure 7:1: Flowchart showing numbers in the dataset we assessed

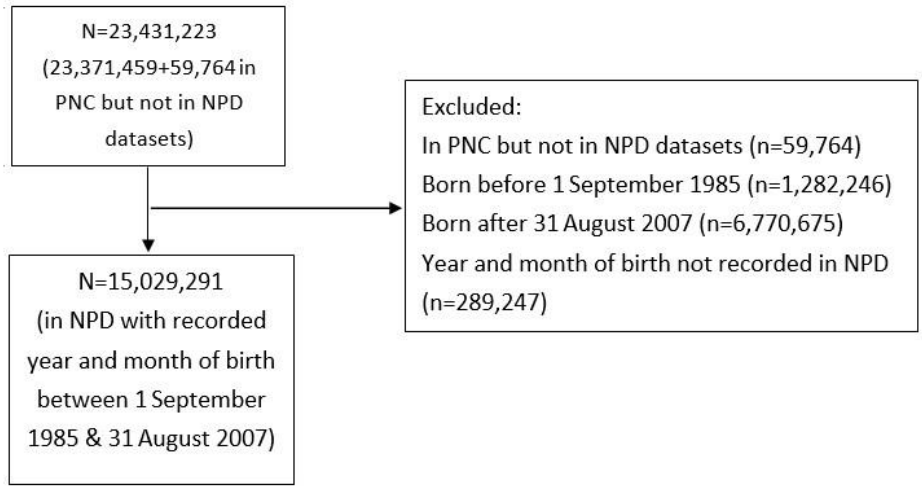


Table 7.1 gives the numbers with at least one year and month of birth in the NPD data and the number of these with at least one record in the PNC (if more than one year and month of birth was recorded, we have used the most frequently occurring one; further details about inconsistencies in year and month of birth are given in Section 10.1). Note that we would expect the percentages to be low in more recent birth cohorts because the offending data only goes up to the end of 2017; thus, for example, those born in 2006/2007 would have only been 10-11 years old at this point.

Table 7.1: Numbers in the linked dataset

Cohort born (academic year)	Total number of individuals with year and month of birth recorded in NPD	Number (%) of these with at least one record in the PNC data
1985/86	653,515	133,689 (20.5%)
1986/87	673,884	139,693 (20.7%)
1987/88	691,920	142,696 (20.6%)
1988/89	685,942	139,341 (20.3%)
1989/90	696,672	138,370 (19.9%)
1990/91	719,977	135,142 (18.8%)
1991/92	716,146	125,354 (17.5%)
1992/93	692,280	108,757 (15.7%)
1993/94	690,696	94,069 (13.6%)
1994/95	677,937	77,334 (11.4%)
1995/96	675,177	60,505 (9.0%)
1996/97	691,482	47,692 (6.9%)
1997/98	686,336	34,886 (5.1%)
1998/99	683,065	24,959 (3.7%)
1999/00	668,269	16,955 (2.5%)
2000/01	652,264	11,744 (1.8%)
2001/02	639,685	7,539 (1.2%)
2002/03	656,386	4,449 (0.7%)
2003/04	675,224	2,094 (0.3%)
2004/05	688,513	714 (0.1%)
2005/06	697,826	177 (<0.1%)
2006/07	716,095	<50
Total	15,029,291	1,446,188 (9.6%)

8 Data availability

8.1 NPD data

In Appendix D we have included tables summarising what education data are present in the linked dataset for each birth cohort, starting with those born in the academic year 1985/96 (thus starting school in 1990/91) and ending with those born in 2006/07 (starting school 2011/12). Tables D1 to D6 detail the school years for which each broad category of data is available (the symbol "X" is used to denote data availability). We have only listed school years reception (age 4/5) to 11 (age 15/16) because, until 2013, the school leaving age was 16 years and, as outlined in Section 6, we only assessed data up to and including Key Stage 4 (up to year 11).

As detailed in Section 2, because the datasets cover different periods of time, particularly in terms of the academic year from which they start, the older cohorts only have limited types of data available on them. For example, the CLA and CiN data start from 2005/06 and 2008/09 respectively, so these data are not available at all for any individuals born before

1989/90 (CLA) and 1991/92 (CiN) and only available for limited school years thereafter, until reaching the cohorts starting school in 2005/06 and 2008/09. Similarly, absence data also started in 2005/06 and school census and exclusions data in 2001/02. So, those born in 1985/86 only have school census data available for year 11 and beyond and only those born from 1997/98 have school census data from reception year onwards.

8.2 PNC data

As indicated above, offenders with at least one PNC or HOCAS record from 2000 onwards and who were born on or after 31st August 1985 were included in the linkage. Here we summarise only the data in the PNC. The dataset contained 16,726,657 records on 9,089,841 offences. Of the 9,089,841 offences, 411,263 (4.5%) were records from 59,764 individuals who had no NPD data in the datasets we requested.

9 Completeness

In Section 7, Figure 7.1 we gave the numbers with missing year and month of birth. The data quality assessments detailed below were carried out among the 15,029,291 individuals who either had a known (certain) year and month birth and were born between 1st September 1985 and 31st August 2007 or those for whom year and month of were uncertain but who had at least one year and month of birth recorded within this range and at least some further NPD data indicating they were born within this range (for example, school census data for a national curriculum year that would indicate they were born within this period).

Overall, levels of completeness were generally high. The main exceptions to this were for ethnicity (Section 9.1.2) and Key Stage 2 and 3 attainment data (Sections 9.2.3 and 9.2.4)

9.1 School census data

The percentages of individuals included in at least one pupil-level school census in each school year ranged from 78% to 88% and were generally slightly higher for the youngest few cohorts. These results are given in Appendix D, Table D7. In the NPD data resource profile (Jay et al. 2019), the figures suggest that between 2011 and 2017 around 6-7% of children aged 5-15 in England were enrolled in an independent school each year (independent schools are not required to return pupil level census data).

Figures also suggest that a relatively small number (estimated to be less than 1%) are home-educated (ADCS, 2019). However, the linked dataset also includes children who would not have been living in England (hence not enrolled in any school in England, state-funded or otherwise) throughout the entire period from reception year to year 11, so we would not expect coverage for any given school year to be equal to the estimated 92-93% who are neither home-educated or enrolled in an independent school. Indeed, the proportion ever appearing in a pupil-level census (including the PRU census and the AP census) for any school year from reception onwards was 92.9%. Among those who had at least one record in the

PNC for an offence classified as triable either way (TEW) or indictment only (IO), this percentage was higher (96.5%).

9.1.1 Year and month of birth

Year and month of birth, gender and ethnicity are recorded in each pupil-level school census. As previously stated, we excluded all those with no recorded year and month of birth as well as those recorded as being born before 1st September 1985 or after 31st August 2007 (unless year and month of birth was uncertain, and they had NPD data indicating they might have been born between 1st September 1985 and 31st August 2007). Thus, our resulting dataset of 15,029,289 individuals only included individuals with at least one year and month of birth recorded. More than one year and/or month of birth was recorded across all the NPD datasets containing this information for 228,144 (1.5%). Further details of these inconsistencies are given in Section 10.

9.1.2 Gender and ethnicity

Of all the individuals born between 1985/86 and 2006/07, 109,461 (0.7%) had no record of gender and 903,027 (5.8%) had no record of ethnicity; a further 12,088 (0.08%) had ethnic group recorded as “refused” and 82,410 (0.5%) as “information not yet obtained”.

9.1.3 Other school census variables: SEN, FSM, IDACI

Aside from the key demographic data, from the pupil-level school census data we assessed three key variables: special education needs (SEN) status, eligibility for free school meals (FSM), and Income Deprivation Affecting Children Index (IDACI) scores, as these indicators are likely to be of interest to most researchers.

Tables D8 to D10 (Appendix D) give the number of non-duplicate records in the main pupil level school census, the Alternative Provision (AP) census and the Pupil Referral Unit (PRU) census (for all school years from reception onwards) for each academic year, and the number of records with missing national curriculum (NC) year, FSM indicator, SEN status, and IDACI score. There was more missing data for IDACI scores than the other two variables.

FSM was complete except in 2001/02 and the numbers missing national curriculum year and SEN status were all very small except from 2001/02 to 2004/05 for SEN status and in 2006/07 and 2007/08 for national curriculum year. For IDACI score, the percentages were all less than 1%, except in the 2001/02 school census (1.2%) and in the 2005/06 summer school census (71.0%) and were slightly higher in the earlier years than the later years (because the numbers are so small for the other variables, percentages are only given for IDACI scores in Table D8). Note that individuals can be included in the school census at more than one school (these pupils are classified as “dual registered”), so the number of records in the school census is greater than the number of individual pupils.

9.2 Attainment

For each key stage, there are many attainment variables. We have assessed variables from early years to Key Stage 4 as listed in Table 9.1. We chose this limited set of variables to give

a broad overview of levels of completeness. Our aim was to choose measures that provided an overall summary of attainment at each key stage but also measures that were included across all years of the data wherever possible, so that it was possible to examine whether completeness had changed over time.

We examined completeness overall as well as separately among individuals who had at least one record in the PNC for an offence classified as triable either way (in magistrates or crown court (TEW)) or indictment only (triable in the crown court (IO)). Where there were apparent differences, we present the data for this subgroup.

For each of the attainment variables we have examined numbers with invalid or missing codes as well as those with a valid entry but a null score; this includes codes such as A (absent), Z (ineligible), X (lost), and others. Note that this is not a data quality issue, but we have examined it because it is relevant to researchers interested in exploring the link between educational attainment and offending.

Table 9.1: Attainment variables examined at each key stage

Key stage	**Measure**	**Variable(s)**
Early years	Personal, social & emotional development score	PSE_total
	Communication, language & literacy score	CLL_total
	Mathematical development score	PSRN_total
	Foundation stage profile total score	EYFSP_total
KS1	Speaking and listening score	KS1_SPEAKANDLISTEN
	Reading score	KS1_READING
	Writing score	KS1_WRITING
	Maths score	KS1_MATHS
KS2	Maths SATs mark	KS2_MATTOTMRK, KS2_MATMRK
	Reading SATs mark	KS2_READMARK, KS2_READMRK
	Writing SATs mark	KS2_WRITMARK, KS2_ENGWRITMRK
KS3	Total English mark	KS3_ENGTOTMRK
	Total maths mark	KS3_MATTOTMRK
KS4	Number of A*- C grades at GCSE/equivalent	KS4_PASS_AC KS4_PASS_AC_PTQ KS4_PASS_AC_PTQ_EE KS4_PASS_AC_3NG_PTQ_EE KS4_PASS_94
	Capped point score (GCSE/equivalents)	KS4_PTSCNEWE KS4_PTSCNEWE_PTQ KS4_PTSCNEWE_PTQ_EE

9.2.1 Early years

The early years attainment data (EYFSP) contained 3,677,455 non-duplicate records (137 were duplicates) for 3,676,234 individuals. 1,121 had more than one record of data once duplicates were removed. Of these, 191 individuals had one earlier record with no valid values; these were removed, leaving 932 individuals with valid data across at least two years

of their early years schooling. Table 9.2 gives numbers and percentages with a valid value but no score (marked as "N") and with an invalid code or value for each of the four variables.

The EYFSP contained data on a 10% subsample until 2006/07, so among those born between 1997/98 and 2000/01 (inclusive), only around 10% had any early years data. Among those with at least some EYFSP data, the percentages with an invalid code or no score were less than 1% except in 1997/98 (the first year of collection) where 1.8% had invalid records.

The proportions with invalid or no scores were similar among those who had a record in the PNC for an offence classified as TEW or IO (data not shown).

Table 9.2: Completeness in EYFSP attainment data: number (%) with invalid and null-scoring data

Cohort born	Number with any data	PSE total		CLL total	
		No score	Invalid	No score	Invalid
1997/98	50,553	<10	901 (1.8%)	<10	890 (1.8%)
1998/99	54,984	98 (0.2%)	<10	97 (0.2%)	<10
1999/00	52,136	<10	88 (0.2%)	<10	105 (0.2%)
2000/01	53,060	79 (0.2%)	<10	81 (0.2%)	<10
2001/02	533,917	599 (0.1%)	67 (<0.1%)	632 (0.1%)	69 (<0.1%)
2002/03	556,395	555 (0.1%)	<10	579 (0.1%)	<10
2003/04	572,701	466 (<0.1%)	24 (<0.1%)	491 (<0.1%)	28 (<0.1%)
2004/05	585,476	489 (<0.1%)	<10	496 (<0.1%)	<10
2005/06	597,520	435 (<0.1%)	<10	476 (<0.1%)	<10
2006/07	619,492	453 (<0.1%)	475 (<0.1%)	471 (<0.1%)	475 (<0.1%)
		PSRN total		**EYFSP total**	
		No score	**Invalid**	**No score**	**Invalid**
1997/98	50,553	<10	908 (1.8%)	<10	<10
1998/99	54,984	101 (0.2%)	<10	94 (0.2%)	16 (<0.1%)
1999/00	52,136	<10	103 (0.2%)	<10	110 (0.2%)
2000/01	53,060	84 (0.2%)	<10	75 (0.1%)	<10
2001/02	533,917	691 (0.1%)	68 (<0.1%)	584 (0.1%)	69 (<0.1%)
2002/03	556,395	637 (0.1%)	<10	534 (<0.1%)	11 (<0.1%)
2003/04	572,701	540 (<0.1%)	27 (<0.1%)	<10	470 (<0.1%)
2004/05	585,476	561 (<0.1%)	<10	477 (<0.1%)	<10
2005/06	597,520	545 (<0.1%)	<10	422 (<0.1%)	<10
2006/07	619,492	528 (<0.1%)	475 (<0.1%)	440 (<0.1%)	475 (<0.1%)

9.2.2 Key Stage 1

In the Key Stage 1 attainment data there were 9,886,117 non-duplicate records (334 were duplicates) for 9,882,880 individuals. Of these, 3,230 individuals had more than one record, of whom 780 had duplicate results (differing by year and/or school). The records for the remaining 2,450 individuals (i.e. records on the same individual with one or more differing value for the four variables) are excluded from Table 9.3.

Valid codes included the different levels: W (working towards level 1) to 6 (individuals with any of these were classified as having a score), as well as A (absent), D (disapplied from national curriculum). The code IN indicated an invalid record. Any other codes were also classified as invalid (there were relatively few of these, but included entries such as "?" or "/" or "-").

As with the early years data, the percentages with an invalid code or no score were all less than 1%. There were no marked differences in the proportions with missing or invalid scores among those with a record for offending (data not shown).

Table 9.3: Completeness in KS1 attainment: number (%) with invalid and null scoring (e.g. absent, disapplied) codes

Cohort born	Number with any data	KS1 speaking and listening		KS1 reading	
		No score	Invalid	No score	Invalid
1990/91	625,351	2,213 (0.4%)	<10	2,194 (0.4%)	<10
1991/92	626,383	2,107 (0.3%)	33 (<0.1%)	2,087 (0.3%)	32 (<0.1%)
1992/93	604,757	2,154 (0.4%)	57 (<0.1%)	2,116 (0.4%)	57 (<0.1%)
1993/94	601,921	2,191 (0.4%)	23 (<0.1%)	2,176 (0.4%)	23 (<0.1%)
1994/95	588,369	2,268 (0.4%)	26 (<0.1%)	2,262 (0.4%)	26 (<0.1%)
1995/96	579,036	2,287 (0.4%)	<10	2,272 (0.4%)	<10
1996/97	588,061	1,870 (0.3%)	<10	1,889 (0.3%)	<10
1997/98	567,853	1,011 (0.2%)	<10	1,044 (0.2%)	<10
1998/99	561,743	995 (0.2%)	<10	983 (0.2%)	<10
1999/00	546,761	719 (0.1%)	<10	736 (0.1%)	<10
2000/01	536,788	732 (0.1%)	<10	698 (0.1%)	<10
2001/02	533,067	553 (0.1%)	<10	572 (0.1%)	<10
2002/03	553,365	593 (0.1%)	<10	613 (0.1%)	<10
2003/04	570,363	654 (0.1%)	<10	621 (0.1%)	<10
2004/05	582,839	697 (0.1%)	<10	684 (0.1%)	<10
2005/06	595,469	875 (0.1%)	<10	851 (0.1%)	<10
2006/07	616,814	867 (0.1%)	<10	785 (0.1%)	<10
		KS1 writing		KS1 maths	
		No score	Invalid	No score	Invalid
1990/91	625,351	2,203 (0.4%)	<10	<10	2,152 (0.3%)
1991/92	626,383	2,094 (0.3%)	32 (<0.1%)	<10	2,131 (0.3%)
1992/93	604,757	2,127 (0.4%)	57 (<0.1%)	<10	2,081 (0.3%)
1993/94	601,921	2,199 (0.4%)	23 (<0.1%)	<10	2,142 (0.4%)
1994/95	588,369	2,279 (0.4%)	26 (<0.1%)	<10	2,240 (0.4%)
1995/96	579,036	2,281 (0.4%)	<10	18 (<0.1%)	2,222 (0.4%)
1996/97	588,061	1,900 (0.3%)	<10	433 (<0.1%)	1,426 (0.2%)
1997/98	567,853	1,043 (0.2%)	<10	1,094 (0.2%)	<10
1998/99	561,743	990 (0.2%)	<10	1,063 (0.2%)	<10
1999/00	546,761	749 (0.1%)	<10	851 (0.2%)	<10
2000/01	536,788	712 (0.1%)	<10	725 (0.1%)	<10
2001/02	533,067	604 (0.1%)	<10	674 (0.1%)	<10
2002/03	553,365	620 (0.1%)	<10	692 (0.1%)	<10
2003/04	570,363	642 (0.1%)	<10	768 (0.1%)	<10
2004/05	582,839	701 (0.1%)	<10	772 (0.1%)	<10
2005/06	595,469	859 (0.1%)	<10	843 (0.1%)	<10
2006/07	616,814	796 (0.1%)	<10	764 (0.1%)	34 (<0.1%)

9.2.3 Key Stage 2

There were two datasets containing Key Stage 2 attainment: KS2Pupil_1996_2019 and KS2Exam_2007_2019. Across both datasets there were 13,108,983 unique individuals; of these, 6,170,992 were only in the pupil file. The first of these datasets (KS2Pupil) contained the variables KS2_READMRK, KS2_ENGWRITMRK, KS2_MATTOTMRK and KS2 _MATMRK and had 13,113, 440 non-duplicate records.

There were 156 duplicate records for our variables of interest (PMR, academic year, and the reading, writing and maths scores). Among these individuals, 13,104, 232 (greater than 99.9%) had only one record in this file, 4,734 had two records, and 17 individuals had three or more records (the maximum was four records). For just under half (41%) of the individuals with more than one record, all records were in the same academic year.

The second of these datasets (KS2Exam) contained the variables KS2_READMARK and KS2_WRITMARK and had 3,144,320 non-duplicate, non-blank records (there were 212 duplicate records for our variables of interest: PMR, academic year, and the reading and writing scores) on 3,144,046 individuals. Among these, 3,143,780 (greater than 99.9%) had only one record and 266 had more than one record (the maximum was three records); for most (91%) of these individuals, all records were in the same academic year and one of these had “absent” in one record, but scores recorded in the other. Taking these out left 99 individuals with more than one record.

Valid values with no score indicated in the metadata were A (absent), M (missing), IN (invalid); further valid codes with no score listed in historical NPD documentation included: B (pupil working below the level of the test), Z (ineligible), X (lost), and F (pupil will take the test in the future) and further codes for invalid included _NV. Those with an entered code of IN or _NV were classified as invalid, as were any other codes not listed in this documentation. The writing scores only went up to the year 2011/12. Table 9.4 gives the number and percentage with either no score or an invalid code.

For those born from 1992/93, the percentage with an invalid code or value was between 4.2% and 4.5% (except those born in 1998/99, when 35% had invalid data and those born in 1992/93, when less than 1% had invalid data), with a similar pattern for reading and writing and maths. The percentages were higher among those born from 1985/86 to 1991/92 and, again, were slightly higher for reading and writing than for maths. Between 1992/93 and 1996/97 up to 1.4% had a null score (absent or other valid code) for the three measures; the numbers in other years were 0 or less than 10.

Table 9.5 gives the numbers and percentages with no score and invalid values for reading among those individuals with a record for an offence (TEW or IO). The percentages with invalid data or null scores were higher than those among all individuals.

Table 9.4: Completeness in KS2 attainment data: number (%) with invalid or null-scoring (e.g. absent) codes - all individuals

Cohort born	Number with any data	KS2 reading		KS2 writing		Number with any data	KS2 maths	
		No score	Invalid	No score	Invalid		No score	Invalid
1985/86	574,660	<10	39,941 (7%)	<10	40,530 (7.1%)	538,312	<10	36,444 (6.8%)
1986/87	608,019	<10	38,169 (6.3%)	<10	38,948 (6.4%)	572,691	<10	35,497 (6.2%)
1987/88	629,830	<10	34,616 (5.5%)	<10	35,371 (5.6%)	599,100	<10	30,881 (5.2%)
1988/89	624,179	<10	33,954 (5.4%)	<10	34,760 (5.6%)	594,732	<10	29,667 (5%)
1989/90	634,276	<10	32,830 (45.2%)	<10	33,463 (5.3%)	605,354	<10	29,173 (4.8%)
1990/91	641,398	<10	32,603 (5.1%)	<10	33,127 (5.2%)	611,972	<10	29,493 (4.8%)
1991/92	637,367	77 (<0.1%)	54,108 (8.5%)	79 (<0.1%)	53,898 (8.5%)	608,851	94 (<0.1%)	28,490 (4.7%)
1992/93	613,154	7,038 (1.1%)	3,694 (0.6%)	7,042 (1.1%)	3,633 (0.6%)	603,538	8,616 (1.4%)	1,085 (0.2%)
1993/94	610,431	4,325 (0.5%)	26,342 (4.3%)	4,258 (0.7%)	26,224 (4.3%)	583,186	5,497 (0.9%)	21,842 (3.7%)
1994/95	595,240	4,014 (0.7%)	24,830 (4.2%)	3,940 (0.7%)	24,826 (4.2%)	569,647	4,678 (0.8%)	20,955 (3.7%)
1995/96	587,601	3,605 (0.6%)	24,889 (4.2%)	3,506 (0.6%)	24,886 (4.2%)	563,232	4,448 (0.8%)	19,877 (3.5%)
1996/97	597,424	4,453 (0.7%)	23,800 (4%)	3,708 (0.6%)	23,768 (4%)	573,044	4,863 (0.8%)	19,502 (3.4%)
1997/98	579,750	<10	25,875 (4.5%)	<10	25,620 (4.4%)	556,852	<10	22,889 (4.1%)
1998/99	571,131	<10	172,274 (30.2%)	<10	172,238 (30.2%)	395,115	<10	176,014 (44.5%)
1999/00	554,739	<10	24,243 (4.4%)	0	24,711 (4.5%)	532,692	<10	22,043 (4.1%)
2000/01	544,044	<10	22,036 (4.1%)	No writing scores from this year		524,439	<10	19,602 (3.7%)
2001/02	540,169	<10	20,094 (3.7%)			521,536	<10	18,623 (3.6%)
2002/03	561,461	<10	19,532 (3.5%)			543,055	<10	18,401 (3.4%)
2003/04	579,261	<10	18,523 (3.2%)			561,826	<10	17,426 (3.1%)
2004/05	592,211	<10	20,999 (3.5%)			572,358	<10	19,836 (3.5%)
2005/06	604,657	<10	21,538 (3.6%)			584,348	<10	20,305 (3.5%)
2006/07	624,188	<10	21,444 (3.4%)			603,358	<10	20,830 (3.5%)

Table 9.5: Completeness in KS2 reading data among those with a record for any offence (TEW/IO)

Cohort born	Number with any data	No score		Invalid	
1985/86	92,214	0		10,598	11.5%
1986/87	99,621	0		10,586	10.6%
1987/88	102,155	0		9,596	9.4%
1988/89	100,199	0		9,096	9.1%
1989/90	98,991	0		8,678	8.8%
1990/91	95,487	<10		8,104	8.5%
1991/92	87,884	<10		8,609	9.8%
1992/93	74,916	1,520	2.0%	348	0.5%
1993/94	64,724	974	1.5%	4,994	7.7%
1994/95	52,662	784	1.5%	3,928	7.5%
1995/96	40,478	620	1.5%	3,229	8.0%
1996/97	31,923	592	1.9%	2,419	7.6%
1997/98	23,271	<10		2,167	9.3%
1998/99	16,910	0		6,331	37.4%
1999/00	11,881	0		1,167	9.8%
2000/01	8,254	0		691	8.4%
2001/02	5,103	0		400	7.8%
2002/03	2,897	0		246	8.5%
2003/04	1,351	0		132	9.8%
2004/05	459	0		63	13.7%
2005/06	105	0		26	24.8%
2006/07	17	0		<10	

9.2.4 Key Stage 3

There were three datasets containing the Key Stage 3 attainment data of interest: KS3Exam_2007_2008, KS3_1998_2006, and KS3Pupil_2007_2008. From 2008/09 to 2012/13 there are variables indicating teacher-assessed levels, but we did not assess these. We have summarised the total point score variables from the maths and English tests from the latter two tables.

KS3Pupil_2007_2008 contained 1,199,894 non-duplicate records for 1,199,521 individuals. 177 individuals had the same scores for the two variables repeated over two years. The records for the remaining 61 individuals (i.e. records on the same individual with one or more differing values for the two variables) are excluded from Table 9.6.

KS3_1998_2006 contained 4,273,236 unique records for 4,266,862 individuals. 4,260,500 (greater than 99%) had only one record, 6,350 individuals had 2 records and 12 individuals had 3 records. In 277 cases, the records were the same over different years, a further 276 individuals had records that were both invalid, and in 3,848 cases one of the records had at least one valid

score and the other was invalid. The remaining 2,238 individuals had valid data across at least two records for different years.

Valid values with no score indicated in the metadata were A (absent) or M (missing). Further valid codes with no score listed in historical NPD documentation included: B and Z (see KS2 section for meaning of these codes). Any other codes were classified as invalid.

Table 9.6 gives the numbers and percentages with no score or an invalid record among all individuals and those with a record for any offence (TEW or IO). For those born between 1985/86 and 1988/89 the percentage with invalid data for English was 9-10% and very few individuals had a missing or absent code; for maths the percentage with invalid data was lower at 5-6%. For those born from 1989/90 to 1993/94, the percentage with invalid data was 4-5% for English (except for those born in 1989/90 when it was less than 1%) and 1-2% for maths. However, the percentage with no score (coded as missing or absent) was 3-5% for English and 2-3% for maths. As such, across all years the percentage with either an invalid or null score ranged between 7.6% and 9.9% for English (except for those born in 1989/90 when it was 5.2%) and between 4.1% and 5.8% for maths. The figures were higher among those with a record for offending (TEW or IO): up to 18.8% for English and 11.0% for maths.

Table 9.6: Completeness in KS3 attainment data: number (%) with invalid and null scoring (e.g. absent) codes

Cohort born	Number with any data	KS3 English No score	Invalid	KS3 maths No score	Invalid
All individuals					
1985/86	584,886	<10	56,911 (9.7%)	<10	34,168 (5.8%)
1986/87	598,464	<10	59,059 (9.9%)	<10	33,857 (5.7%)
1987/88	614,821	21 (<0.1%)	55,342 (9.0%)	18 (<0.1%)	33,489 (5.5%)
1988/89	607,781	494 (<0.1%)	57,897 (9.5%)	312 (<0.1%)	29,852 (4.9%)
1989/90	616,073	30,756 (5.0%)	1,158 (0.2%)	16,661 (2.7%)	9,336 (1.5%)
1990/91	621,115	24,470 (3.9%)	29,970 (4.8%)	18,715 (3.0%)	10,819 (1.7%)
1991/92	618,336	22,155 (3.6%)	28,944 (4.7%)	16,634 (2.7%)	9,909 (1.6%)
1992/93	596,180	20,446 (3.4%)	29,034 (4.9%)	15,667 (2.6%)	9,535 (1.6%)
1993/94	592,334	19,705 (3.3%)	25,618 (4.3%)	14,927 (2.5%)	9,382 (1.6%)
Those with at least one record in the PNC for any offence (TEW/IO)					
1985/86	95,016	0	17,610 (18.5%)	0	10,409 (11.0%)
1986/87	99,476	0	18,731 (18.8%)	0	10,960 (11.0%)
1987/88	100,630	0	18,053 (18.0%)	0	10,989 (10.9%)
1988/89	98,205	74 (<0.1%)	18,511 (18.8%)	71 (<0.1%)	9,892 (10.1%)
1989/90	97,061	10,332 (10.6%)	157 (1.2%)	6,450 (6.6%)	2,649 (2.7%)
1990/91	93,242	8,896 (9.5%)	7,627 (8.2%)	7,464 (8.0%)	1,802 (1.9%)
1991/92	86,120	8,118 (9.4%)	6,946 (8.1%)	6,636 (7.7%)	1,556 (1.8%)
1992/93	73,514	7,090 (9.6%)	6,258 (8.5%)	5,895 (8.0%)	1,511 (2.1%)
1993/94	63,515	5,841 (9.2%)	5,083 (8.0%)	5,115 (8.1%)	1,326 (2.1%)

9.2.5 Key Stage 4

There were four datasets containing Key Stage 4 attainment data: KS4Pupil_2002_to_2014 contained the variables KS4_PASS_AC, KS4_PASS_AC_PTQ, KS4_PTSCNEWE and KS4_PTSCNEWE_PTQ and had 8,362, 282 unique records for 8,277,060 individuals. The majority of individuals (8,193, 225, 99%) had only one record, 81,458 had 2 records, 1,367 had 3 records and 10 had 4 records. Of these 83,835 individuals with more than one record, 8,692 had the same values for the attainment variables in each record. A further 16,893 individuals with valid data had second record containing null data, leaving 58,250 with more than one non-duplicate record, where both records contained at least one variable of differing valid data. This would be expected, as individuals can sit their GCSEs (or equivalent exams) early, and individuals might retake exams.

KS4Pupil_2015_to_2019 contained the equivalent variables for later years: KS4_PASS_AC_PTQ_EE, KS4_PASS_AC_3NG_PTQ_EE, KS4_PASS_94 and KS4_PTSCNEWE_PTQ_EE. There were 3,023,063 non-duplicate records for 3,007,476 individuals. Among these, 2,991,946 (greater than 99.5%) had a single record, 4,883 individuals had the same scores over two or more years, 3,098 had one valid record, where at least one other was blank; this left 7,549 with at least one valid value in each of their records.

Among all individuals, 4,141 had a record in both files. Thus, the total number of individuals with a record in either of the files was 11,280,395. Of these, 2,985 had a record for both the earlier and one of the later measures of number of passes at grade A* to C and 1,886 had a record for both the earlier and later measure of the capped point score. Altogether, 11,279,266 had a valid record for number of A*-C grades (and their equivalents) and 11,278,509 had a valid value for the capped point score. The proportions with invalid data were small (the maximum was 1.7%); these are shown in Table 9.7. The proportions with invalid data for the capped point score were higher among those with an offending record (Table 9.7).

Table 9.7: Completeness in KS4 attainment data: number (%) with missing or invalid data among all individuals and those with a record for any offence (TEW/IO)

	All individuals			Ever offended (TEW/IO)	
Cohort born	**Number with any data**	**Passes A* to C**	**Capped point score**	**Number with any data**	**Capped point score[1]**
1985/86	589,530	0	Not measured		
1986/87	621,829	13 (<0.1%)	Not measured		
1987/88	647,443	0	23 (<0.1%)		
1988/89	643,347	0	<10		
1989/90	652,748	0	<10		
1990/91	661,098	0	461 (<0.1%)	92,316	129 (0.1%)
1991/92	661,671	0	11,220 (1.7%)	85,805	4,686 (5.5%)
1992/93	640,892	0	8,286 (1.3%)	73,780	3,134 (4.3%)
1993/94	644,105	0	6,209 (1.0%)	65,683	1,997 (3.0%)
1994/95	631,992	0	10 (<0.1%)		
1995/96	625,769	0	<10		
1996/97	638,260	0	<10		
1997/98	623,952	0	<10		
1998/99	618,246	0	<10		
1999/00	604,266	0	<10		
2000/01	588,634	0	<10		
2001/02	584,990	0	<10		
2002/03	600,722	0	<10		

1. There were fewer than 10 individuals with missing/invalid results for all years other than those shown and fewer than 10 individuals with missing/invalid data for number of A* to C passes in all years.

9.3 Absence

Absence is recorded as authorised on unauthorised. From the data available, it is possible to calculate, for each individual and each academic year, the percentage of sessions (half days) missed overall and – separately - due to authorised and unauthorised absence. The percentage in each school year for whom percentage absence was calculable (among those with school census data in that year) ranged from 96.9% to 99.7%; the majority were over 99%. These results are given in Appendix D, Table D11.

9.4 CLA/CiN

In this section we have summarised data from the full episode CLA data and the CLA data included in the NPD; as already stated, the full episode data is not included in the linked dataset as standard.

In the full episode CLA data, there was data on 868,138 non-duplicate CLA episodes (32 were duplicate records) for 244,342 individuals. In the CLA data included in the NPD, there was information on an additional 65,228 episodes; these data included 12,904 individuals who did not appear in the full episode data. Most key variables had no missing information; the reason the episode ended was missing for 0.1% of episodes.

The CiN data contained 6,576,678 records up to academic year 2017/2018 (inclusive) on 2,074,450 individuals. These records related to 3,991,295 CiN episodes. Among these episodes, primary need was not recorded for 181,063 (4.5%); a further 264,658 (6.6%) episodes had primary need recorded as not stated or unknown code. Referral source was missing for 1,990,658 (49.9%); a further 81,652 (2.0%) episodes had source unknown or invalid. The last record in a given episode was assumed to be closed for 3,228,376 (CiN=0 on 31st March). Reason for closure was missing or unknown/not stated/invalid for 306,852 (9.5%) of these.

9.5 PNC data

9.5.1 Offence codes

Among the 9,089,841 offences, there were 714,434 (7.9%) with Home Office offence codes not listed in the offence codes table included in the metadata. These are presented by type in Table 9.8. A large proportion (86%) were breach offence codes and most (94,421) of the remaining ones were codes from outside England; 4,046 (0.6% of those with unlisted codes; 0.04% of all offences) were invalid codes.

The 94,421 offences (1% of all the offences) with codes from Scotland, Northern Ireland, Jersey, Guernsey, and the Isle of Man were from 19,481 individuals. Of these, 11,962 (61.4%) did not have any NPD data in the datasets we requested.

Table 9.8: Unlisted Home Office offence codes

Type of code	Number of offences
Breach code	615,967
Scotland	65,175
Northern Ireland	17,571
Jersey	4,271
Guernsey	2,905
Isle of Man	4,499
Other: missing or invalid	4,046
Total	714,434

9.5.2 Offence dates

There were 65,097 offences (0.7% of all offences) with unknown offence date (recorded as 1 January 1900 in the data). Of these 64,504 (99.1%) had a Scottish offence code. Of the remaining 9,024,744 offences, 9,024,128 (greater than 99.9%) had an offence date between 1995 and 2017 (inclusive). Among the 616 with a recorded offence date outside this range, most were obvious errors (for example, 2030 recorded instead of 2003), as fewer than thirty also had a court or caution date outside this range.

9.5.3 Age at offence

Among all the offences (n=9,089,841), 65,097 (0.7%) had no age at offence recorded; these were the offences with unknown offence data. A further 543 had an age that was zero or negative and 1,578 (less than 0.1%) had an age that was under 10 years.

9.5.4 Adjudication and disposal codes

All offences had a recorded adjudication code. The distribution of these is shown in Table 9.9. Of the offences where the outcome was guilty, not guilty or non-conviction, all except 36 (less than 0.01%) had a disposal code recorded.

Table 9.9: Adjudication codes

Adjudication code[1]	**Number (%) of offences**
G: Guilty	5,165,914 (57%)
N: Not guilty	809,945 (9%)
J: Caution/warning/reprimand guilty	2,049,226 (23%)
O: Non-conviction	1,057,473 (12%)
P: Pending	7,283 (<1%)
All	9,089,841

1. Outcomes N, O and P are only held by the MoJ in specific circumstances, including where they are accompanied by a conviction or subject to later update.

10 Consistency and uniqueness

10.1 Year and month of birth

Among the 15,029,289 individuals with year and month of birth recorded as being between 1st September 1985 and 31st August 2007, more than one year and/or month of birth was recorded (across all the NPD datasets containing this information) for 228,144 (1.5%), 224,911 had two recorded, 3,164 had three and 69 had four or more recorded. Among these, the year and month of birth corresponded to the same academic year of birth (for example born 1985/86, or 1st September 1985 – 31st August 1986) for 160,836 (71%); the remaining 67,308 had year/months of birth corresponding to more than one academic year of birth. Of these, 59,838 (89%) had corresponding academic years a maximum of one year apart (i.e. either had the same calendar year of birth recorded each time and only month of birth differed or both year and month of birth differed).

For most individuals with corresponding academic years more than one year apart, entries in the NPD for national curriculum year could be used to determine their correct (presumed) academic year of birth or the second year and month of birth was almost certainly a recording error (e.g. as it appeared only once).

Individuals with at least one record in the PNC were more likely to have more than one year and month of birth recorded (2.2% of individuals with a record in the PNC compared to 1.4% of individuals with no record in the PNC).

10.1.1 Consistency in year of birth between the NPD and PNC

The PNC data did not contain year or month of birth but did include age at offence. From this, we could calculate a derived year of birth (except when age was missing). For those 1,446,188 individuals in the PNC and in our final 15,029,291, at least one age was recorded for 1,441,919 (99.7%). For a small minority of these (642 individuals), their recorded ages resulted in more than one year of birth. For the remaining 1,441,277 the year of birth recorded in the NPD was incompatible with the derived year of birth in the PNC for 4,885 (0.3%) individuals. For 2,948 (60%) of these, the year of birth derived from the PNC was only one or two years earlier than the year of birth recorded in the NPD.

For the remainder, the year of birth derived from the age at offence recorded in the PNC was up to 22 years earlier or up to 18 years later than the year of birth recorded in the NPD, but these large differences only applied to a very small number of individuals and, among the 516 where the year of birth was different by more than 5 years around two thirds (324) were different by 10 years, which suggests likely recording errors (for example, some of these individuals had offences at very young ages of less than 10 years, according to their PNC-derived year of birth).

10.2 Gender

Among the individuals with at least one recorded value for gender, 14,736,075 (98.8%) had only one gender recorded within the different datasets of the NPD. Those with a record in the PNC were slightly less likely to have more than one gender recorded (1.1% - compared to 1.3% among those with no record in the PNC). Note that this is not necessarily a data quality issue because the school census collects gender according to the wishes of the pupil and/or parent and gender identity can change.

10.3 Ethnicity

Within the school census from 2004, ethnicity is either parent or student reported. Among the 14,502,553 individuals with at least one recorded value for ethnicity in the DfE data, 2,086,373 (14.4%) had more than one minor ethnic group recorded. This has nineteen categories excluding refused and invalid: Bangladeshi, Indian, Pakistani, Chinese, Any Other Asian Background, Black African, Black Caribbean, Any Other Black Background, White British, White Irish, Traveller of Irish Heritage, Gypsy/Roma, Any Other White Background, White and Asian, White and Black African, White and Black Caribbean, Any Other Mixed Background, Any Other Ethnic Group. 709,351 (4.9%) individuals had more than one major ethnic group recorded. This has six categories, which are derived from the minor groups, excluding unclassified: Asian, Black, Chinese, Mixed, White, Any Other Ethnic Group. As with gender, this is not necessarily a data quality issue since an individual can change their ethnic perception and how they wish their ethnicity to be recorded.

10.4 School census data

Across all school census datasets, a small proportion (mostly less than 1%; up to 1.7% in a minority) of individuals appeared in different academic years with the same national curriculum year (for example, year 1 in 2005/2006 and year 1 in 2008/2009). The numbers of individuals this relates to are shown in Appendix D, Table D12.

Overall, the proportion with at least one instance of the same national curriculum year appearing in two different academic years (across all school years from reception to year 11) ranged from 0.2% (those born 1985/86) to 3.6% (those born 1995/96 and 1996/97). The proportions were generally higher among those who had at least one record for an offence (TEW or IO) in the PNC (4.7% for those born in 1996/97). These figures are shown in Table 10.1.

In addition, a small number of individuals appeared in the same school census with two different national curriculum years recorded. These numbers are given in Appendix D, Table D13.

Table 10.1: Numbers (%) with at least one instance of the same national curriculum year appearing in two different academic years across all school years

Cohort born	All individuals		Those with record for any offence (TEW/IO)	
	Number in at least one school census	Number (%) with same NC year in >1 academic year	Number in at least one school census	Number (%) with same NC year in >1 academic year
1985/86	555,183	1,279 (0.2%)	86,309	114 (0.1%)
1986/87	590,363	9,320 (1.6%)	98,289	1,433 (1.5%)
1987/88	616,904	12,101 (2.0%)	103,064	2,026 (2.0%)
1988/89	616,170	11,808 (1.9%)	102,063	2,165 (2.1%)
1989/90	630,774	12,857 (2.0%)	101,696	2,310 (2.3%)
1990/91	654,399	9,569 (1.5%)	99,105	1,837 (1.9%)
1991/92	656,819	19,679 (3.0%)	91,437	3,154 (3.5%)
1992/93	640,626	18,317 (2.9%)	78,581	2,547 (3.2%)
1993/94	645,408	19,177 (3.0%)	68,016	2,413 (3.6%)
1994/95	637,969	20,668 (3.2%)	55,479	2,314 (4.2%)
1995/96	635,176	22,773 (3.6%)	42,817	1,954 (4.6%)
1996/97	651,104	23,140 (3.6%)	33,764	1,576 (4.7%)
1997/98	637,918	17,698 (2.8%)	24,748	1,022 (4.1%)
1998/99	630,884	15,793 (2.5%)	17,989	729 (4.1%)
1999/00	617,239	16,976 (2.8%)	12,688	565 (4.5%)
2000/01	607,192	10,611 (1.8%)	8,751	361 (4.1%)
2001/02	602,198	7,448 (1.2%)	5,429	203 (3.7%)
2002/03	621,708	7,552 (1.2%)	3,083	120 (3.9%)
2003/04	635,066	8,331 (1.3%)	1,446	59 (4.1%)
2004/05	643,950	5,927 (0.9%)	489	17 (3.5%)
2005/06	653,590	5,814 (0.9%)	Numbers suppressed	
2006/07	667,042	4,783 (0.7%)	Numbers suppressed	

11 Summary of findings

The work carried out for this report formed the first stage of a wider project carried out to assess the feasibility of using the MoJ-DfE linked dataset for evaluating early interventions for violence prevention. The aim of this first stage of the project was to assess the reliability of the data included in the linked MoJ-DfE dataset. However, it is important to point out that the data we were granted access to (both in terms of datasets and in terms of variables) were those needed to carry out the feasibility study and not all datasets included in the data share (or all variables within them).

In this report, we have given a brief overview of the datasets included in the linked dataset. We have then reported, firstly, on the linkage process, using the GUILD guideline as a basis for our assessments and, secondly, on quality assessments carried out on a subset of variables (from a subset of datasets) in the linked dataset.

Our assessment of the linkage process relies on information provided to us by the MoJ and DfE. The main findings from this are summarised below:

In February 2020, the MoJ sent the DfE 2,631,842 internally linked identifiers (IDs) from their police and magistrates court data systems. The DfE ran iterative deterministic matches against the identifiers from the education records. They used increasingly less restrictive versions of the individual identifiers, with 103 steps divided into 7 descending matching priorities. A separate team within the DfE performed a manual clerical review of all the cases that were not matched on the full exact set of identifiers. All the MoJ IDs, both matched 1,917,825 (73%) and unmatched 714,017 (27%), were then transferred back to the MoJ.

The MoJ data team reported the aggregate proportions of matched and unmatched records by key attributes. They then removed 29,944 (1.1%) cases where many IDs had linked to many alternative IDs, leaving 45,403 cases. A further 40,324 cases (1.5% of the original cases) were removed, as they were not the best match available, based on the source files used for matching. The remaining 5,079 of removed individuals were cases where the source files were equal, but the matches were different. The final MoJ dataset included 1,842,478 individuals (70% of the original 2,631,842). The MoJ data team reported the aggregate proportions of retained and removed individuals by key attributes.

Linkage error arises in two ways: false matches (matched records that in fact come from different individuals) and missed matches (two records that belong to the same individual but are not matched). In order to evaluate likely linkage error and – as a result – the potential impact of this on research findings, it is firstly necessary to understand the details of the linkage process and, secondly, to be able to assess whether the unmatched individuals differ systematically from the matched individuals.

The MoJ have written a report (MoJ, 2021b) providing aggregate characteristics of matched and unmatched individuals, both before and after the removal of records. We were able to use information from that, some of which is included in our report, to make some assessments of differences between these groups. In the MoJ-DfE linked dataset, a certain proportion of the unmatched individuals will be people resident in Wales (the MoJ datasets cover England and Wales, whereas the NPD covers England only). In the magistrates' court data, Welsh police forces accounted for 36% of the unmatched records (this was offences not individuals). Conversely, 2% of the matched records were from Welsh police forces. Recent MoJ analyses of the home postcodes in the PNC showed 9% of English postcodes remained unmatched to the NPD. Conversely 78% of the Welsh postcodes were unmatched (MoJ, 2021b).

There were also a small number of police records with out-of-England codes. However, since we did not have access to any data on the unmatched cases, including indicators which could indicate higher missed matches in vulnerable or marginalised population sub-groups (such as no fixed abode as an indicator of homelessness), we cannot provide a breakdown of characteristics of the remaining unmatched individuals or a comparison of whether or how these individuals differ from the matched individuals.

Similarly, without data on unmatched cases, and because we did not have access to details of all the processes involved in the linkage, particularly the cleaning processes and the clerical review, we cannot comment on the likely impact of these processes of there being missed or false matches.

The match quality dataset indicated that the matching was more certain when the MoJ identifiers came from the PNC and the DfE identifiers came from the school census data. Researchers can use this information to carry out sensitivity analyses. Further, MoJ will use only the PNC data source in future NPD linkage projects.

In our data quality assessments we focussed mainly on completeness, consistency and uniqueness of key variables. Where appropriate, we made assessments among those with an offending record as well as among all individuals in the dataset. From the MoJ datasets, we assessed variables in the PNC; from the DfE, we assessed key variables in a number of different datasets. It is important to point out that in any large administrative dataset, some degree of incompleteness and inconsistency is inevitable.

There was data on over 23 million individuals in the NPD datasets extracted for this project. Approximately 1% of these had missing information on year and month of birth and there were just over 15 million individuals born between September 1985 and August 2007; we restricted our assessments to these individuals. Among these, 7.1% did not appear in any school census from reception year onwards. This would include individuals who were educated in an independent school or at home for this whole period and anyone who lived outside England for this whole period.

We found that levels of completeness in most of the key variables we assessed were high. This was mainly for variables from the NPD but also included key variables from the PNC (Home Office offence code, offence date, age at offence, adjudication code and disposal codes). The main exceptions to this were ethnicity and the Key Stage 2 and Key Stage 3 attainment variables. Just over 6% of individuals had no ethnicity recorded or ethnicity classified as refused or not yet obtained. A further 14% had more than one minor ethnic group recorded (this corresponded to 5% having more than one major ethnic group recorded). In terms of attainment, up to 8% in Key Stage 2 and up to 10% in Key Stage 3 had invalid data for the attainment variable of interest. These percentages were higher among those individuals who had a record in the PNC for any offence classified as triable either way or indictment only.

Levels of inconsistency / differences over time were generally very low. Ethnicity was possibly the most affected variable in this respect although, as described above, this may not necessarily be a data quality issue in this instance. There was evidence that certain inconsistencies (for example, in recorded National Curriculum year) were slightly higher among offenders.

Across the two sources of data, we were able to make some assessment of potential inconsistencies in terms of year of birth. The MoJ data included individuals born on or after 31st August 1985 who had at least one record in the PNC from 2000 onwards. The DfE data we were provided with contained data from individuals born prior to this. There was a small number of individuals (less than 1000) who had a record in the PNC but whose year and month of birth recorded in the NPD was prior to August 1985. Further, we were able to carry out an approximate comparison of year of birth, since the DfE data includes year and month of birth and the PNC data contains age at offence. We found that over 99% of individuals had ages that were compatible between the two data sources and, for the majority of individuals where there was potential disagreement, this was relatively small (one or two years' difference).

Finally, we found that data quality in the NPD datasets generally improved over time.

There are limitations to what we have been able to do. Firstly, we did not have access to full details of the data linkage process. As such, we can only make general observations about specific aspects of this.

Our assessment of the data was necessarily restricted to the subset of data needed to address the main aim of our project. In particular, we did not have access to the following MoJ datasets: CREST, prison receptions, prison discharges, prison population and OASys. Further, for practical reasons, we could not assess all variables in all datasets included in our extract. Within the datasets included in our extract, we restricted our evaluations to what we regarded as key variables likely to be used by researchers, primarily for the period from reception (age 4/5) up to the end of Key Stage 4 (age 16).

In summary, the quality of the data that we assessed in the linked dataset was high, particularly bearing in mind that the data originate from large administrative data sources that do not share a common unique identifier. Since some parts of the linkage process were undocumented and because we did not have access to all documentation about the process, it is difficult for us to draw conclusions about linkage error and we therefore cannot make any comments on the likely impact of this on the data.

Our main recommendation in terms of improving the value of the dataset for research purposes would be for researchers to be granted access to pseudonymised data on key characteristics as recorded in the justice data for unmatched individuals such as year of birth, ethnicity, gender, and limited offending data such as age at first offence, number of offences, and other summary information. This would allow them to make their own assessments of potential biases which could be tailored to the specific purpose of their research. There is also value in providing the same attributes from the Justice data for those individuals who are linked from sources where their attributes are unavailable.

Appendix A: References

Association of Directors of Children's Services Ltd. (2019). Elective Home Education Survey 2019. Manchester: England. Retrieved from: https://adcs.org.uk/assets/documentation/ADCS_Elective_Home_Education_Survey_Analysis_FINAL.pdf

Department for Education. (2010). OSR24/2010 Statistical Release. Referrals, assessments and children who were the subject of a child protection plan (children in need census – provisional) year ending 31 March 2010. London. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/219409/osr24-2010.pdf

Department for Education. (2017). A guide to exclusion statistics, DFE-00226-2015. Crown: London. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/642577/Guide-to-exclusion-statistics-05092017.pdf.

Department for Education. (2019a). A guide to absence statistics, DFE-00148-2015. Crown: London. Retrieved from: https://www.gov.uk/government/publications/absence-statistics-guide.

Department for Education. (2019b). Children looked after by local authorities in England: technical specification to the 2020 to 2021 data collection, DfE-00215-2019. Crown: London. Retrieved from: https://www.gov.uk/government/publications/children-looked-after-return-2020-to-2021-technical-specifications.

Department for Education. (2021). School census 2020 to 2021: technical information, DfE-00060-2020. Crown: London. Retrieved from: https://www.gov.uk/government/publications/school-census-2020-to-2021-technical-information

Education Standards Analysis and Research Division. (2011). Research Report DFE-RR171: A profile of pupil absence in England. Department for Education: London. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/183445/DFE-RR171.pdf.

Emmott, E.H., Jay, M.A., Woodman, J. (2019). Cohort profile: Children in Need Census (CIN) records of children referred for social care support in England. *BMJ Open 2019*; 9:e023771. doi:10.1136/bmjopen-2018-023771.

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L.-C., Smith, P., Dibben, C., Goldstein, H. (2018). GUILD: GUidance for Information about Linking Data sets. *Journal of public health (Oxford, England),* 40(1):191–8. doi:10.1093/pubmed/fdx037.

Jay, M.A., McGrath-Lone, L., Gilbert, R. (2019). Data Resource: the National Pupil Database (NPD). *International Journal of Population Data Science,* 4(1). doi:10.23889/ijpds.v4i1.1101.

McGrath-Lone, L., Harron, K., Dearden, L., Nasim, B., & Gilbert, R. (2016). Data Resource Profile: Children Looked After Return (CLA). International Journal of Epidemiology, 45(3):716-17. doi:10.1093/ije/dyw117.

Ministry of Justice. (2011). (2019). Guide to criminal justice statistics. Crown: London.

Ministry of Justice. (2020). Guide to criminal court statistics. Crown: London.

Ministry of Justice. (2021) Data First: An Introductory User Guide. Crown: London. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/984510/data-first-user-guide-v5.pdf

Ministry of Justice. (2021b). MoJ-DfE dataset: Bias Testing Report. Crown: London (unpublished).

The National Archives. (2021, April). UK Government Web Archive. Retrieved from: http://www.nationalarchives.gov.uk/webarchive/

Appendix B: List of variables requested

Dataset	Variables
Anon Unique IDs	MoJuid Dataid PupilMatchingRefAnonymous MatchRuleApplied
Match quality dataset	MoJUID PupilMatchingRefAnonymous Top_rule MoJ_source DfE_source PMR_selection_flag UID_selection_flag
DfE datasets	
SC_Pupil_01-02_to_17-18	PupilMatchingRefAnonymous YearOfBirth MonthOfBirth Ethnicity EthnicGroup EthnicGroupMinor EthnicGroupMajor FSMeligible FirstLanguage MotherTongue Language LanguageGroup LanguageGroupMinor LanguageGroupMajor EntryDate DateOfJoining LeavingDate DateOfLeavingSch SENprovision SENstage SENstatus PrimarySENtype SecondarySENtype MobilityInd COA LSOA01 LLSOA_[term[yy]) LSOA11 IDACIScore LA LA_9Code URN Gender NCyearActual ActualNCYearGroup TypeOfClass NurseryClassInd ClassType
PRU_Census_09-10_to_12-13	PRU_YearOfBirth PRU_MonthOfBirth PRU_Ethnicity PRU_EthnicGroupMinor PRU_EthnicGroupMajor PRU_FSMeligible PRU_Language PRU_LanguageGroupMinor PRU_LanguageGroupMajor

	PRU_EntryDate PRU_SENprovision PRU_SENprovisionMajor PRU_PrimarySENtype PRU_SecondarySENtype PRU_OACODE PRU_LLSOA PRU_IDACISCORE PRU_LA PRU_LA_9Code PRU_URN PRU_Gender PRU_PupilMatchingRefAnonymous PRU_NCyearActual
EarlyYearsCensus_07-08_to_12-13	EYC_YearOfBirth EYC_MonthOfBirth EYC_Ethnic_Code EYC_EthnicGroupMinor EYC_SEN_provision EYC_LA EYC_URN EYC_PupilMatchingRefAnonymous EYC_ACADYR
Alt_Provision_07-08_to_17-18	AP_YearOfBirth AP_MonthOfBirth AP_Ethnicity AP_EthnicGroupMinor AP_EthnicGroupMajor AP_FSMeligibility AP_FSMeligible AP_SENProvision AP_PrimarySENtype AP_SecondarySENtype AP_LA AP_ACADYR AP_PupilMatchingRefAnonymous AP_Gender
EYFSP_02-03_to_12-13	FSP_ACADYR FSP_PupilMatchingRefAnonymous FSP_GENDER FSP_NFTYPE FSP_PSE_TOTAL FSP_CLL_TOTAL FSP_PSRN_TOTAL FSP_RKUW FSP_RIPD FSP_RICD FSP_EYFSP_TOTAL
KS1_97-98_to_14-15	KS1_ACADYR KS1_PupilMatchingRefAnonymous KS1_RECORD_STATUS KS1_GENDER KS1_ToE_CODE KS1_NFTYPE KS1_MOB1 KS1_MOB2 KS1_TRIALFLAG KS1_SPEAKANDLISTEN KS1_READING KS1_WRITING KS1_MATHS KS1_SCIENCE

KS2_00-01_to_17-18	KS2_ACADYR KS2_PupilMatchingRefAnonymous KS2_GENDER KS2_ToE_CODE KS2_NFTYPE KS2_NPDDEN_NAT KS2_READMRK KS2_ENGWRITMRK KS2_ENGTOTMRK KS2_MATTOTMRK KS2_SCITOTMRK KS2_SCIPOINTS KS2_READMARK KS2_WRITMARK
KS3_03-04_to_12-13	KS3_ACADYR KS3_PUPILMATCHINGREFANONYMOUS KS3_GENDER KS3_TOE_CODE KS3_NFTYPE KS3_MOB1 KS3_MOB2 KS3_LEVLENGTA KS3_LEVLMATTA KS3_LEVLSCITA KS3_LEVXENGTA KS3_LEVXMATTA KS3_LEVXSCITA KS3_LEVAXENGTA KS3_LEVAXMATTA KS3_LEVAXSCITA KS3_ENG1MRK KS3_ENG2MRK KS3_ENGREADMRK KS3_ENGWRTTOTMRK KS3_ENGTOTMRK KS3_MATPAP1MRK KS3_MATPAP2MRK KS3_MATARTHMRK KS3_MATTOTMRK KS3_SCIPAP1MRK KS3_SCIPAP2MRK KS3_SCITOTMRK
KS4_05-06_17_18	KS4_ACADYR KS4_PupilMatchingRefAnonymous KS4_GENDER KS4_ToE_CODE KS4_NFTYPE KS4_NEW_TYPE KS4_NEWER_TYPE KS4_STATUS KS4_PASS_AC KS4_PASS_AG KS4_PTSTNEWE KS4_PTSCNEWE KS4_PTSTNEWG KS4_PTSCNEWG KS4_POINTS_OLD_G KS4_PTSCOLDG KS4_LEVEL2_EM KS4_PASS_AC_PTQ_EE KS4_PASS_AC_3NG_PTQ_EE KS4_PASS_94 KS4_PASS_LEV2_PTQ_EE KS4_PASS_LEV2EM_PTQ_EE

	KS4_PASS_AG_PTQ_EE KS4_PASS_LEV1_PTQ_EE KS4_PTSTNEWE_PTQ_EE KS4_PTSCNEWE_PTQ_EE
KS5_06-07_to_17-18	KS5_URN KS5_ACADYR KS5_PupilMatchingRefAnonymous KS5_Gender KS5_TOTPTS KS5_TOTPTSE
CIN_08-09_to_16-17	CIN_CINAt31March CIN_PupilMatchingRefAnonymous CIN_LatestReferralDate CIN_LatestClosureDate CIN_AnyCasesOpen31March CIN_CPPstartDate CIN_CPPendDate CIN_CategoryOfAbuse CIN_InitialCategoryOfAbuse CIN_LatestCategoryOfAbuse CIN_NumberOfPreviousCPP CIN_ACADYR CIN_CINReferralDate CIN_ReferralSource CIN_PrimaryNeedCode CIN_CINClosureDate CIN_ReasonForClosure CIN_DateOfInitialCPC CIN_ReferralNFA CIN_CaseOpen31March CIN_AgeStartOfCINPeriod CIN_AgeEndOfCINPeriod CIN_ACADYR CIN_Disability
CLA_05-06_to_16-17	cla_PupilMatchingRefAnonymous cla_CLA_LA cla_PROCESSING_YEAR cla_UASC_STATUS cla_POC_START cla_POC_LENGTH cla_CAT_NEED cla_EPI_IDX cla_DATE_EPI_COMM cla_DATE_EPI_CEASED cla_LEGAL STATUS cla_PLACEMENT cla_REC cla_CLA_31_MARCH cla_CLA_6_MONTHS cla_CLA_12_MONTHS cla.CLA_PP_6_MONTHS cla.CLA_PP_1_DAY cla.REASON_PLACE_CHANGE
Absence_05-06_to_17-18	SourceTable_1_ab SourceTable_2_ab SourceTable_3_ab OnRoll_1 OnRoll_2_ab OnRoll_3_ab AcademicYear_ab PupilMatchingRefAnonymous_ab Gender_ab IDACI_S_school IDACI_R_school

	SessionsPossible_3Term_ab AuthorisedAbsence_3Term_ab UnauthorisedAbsence_3Term_ab MainRecord_ab
Exclusions_01-02_to_04-05	PupilMatchingRefAnonymous_ex Gender_ex StartDate_ex PermanentExclusionInd_ex
Exclusions_05-06_to_16-17	Reason_ex AcademicYear_ex PupilMatchingRefAnonymous_ex Category_ex StartDate_ex Sessions_ex TotalFixedExclusions_ex TotalFixedSessions_ex PermanentExclusionIndicator_ex PermanentExclusionCount_ex Category_[n]_ex Sessions_[n]_ex NCYear_[n]_ex
NCCIS_10-11_to_16-17	NCCIS_ACADYR NCCIS_s.PupilMatchingRefAnonymous NCCIS_Current_Activity_Code NCCIS_NEET_Start_Date NCCIS_Year_11_Intended_Destination NCCIS_Year_11_Offer NCCIS_Year_12_Offer NCCIS_Y11_September_Guarantee_status NCCIS_Y12_September_Guarantee_status
Linked CLA Return (SSDA903)	All variables
School-level census data	School-level data on absence, %FSM, attainment, and others
MoJ datasets	
PNC	MoJUID CaseID OffenceID CourtCautionDate Condition Text EndDate MaxAdjudicationCode Cautiontype DisposalID PNCDisposalCode HODisposalCode DisposalDuration ACPOCode CCCJSCode HOOffenceCode OffenceStartDate OffenceStationCode AdjudicationCode OffenceStartAge Cautiontype
HOCAS[1]	MoJUID OFFENCE_ID OFFENDER_TYPE OFFENCE_DATE ARREST_DATE FINDING

1. This dataset was not included in our extract because Covid-19 prevented access

Appendix C: Match priority table

Match no.	Priority	Data source	Rule	Comment	No. of matches (records)	No. of matches (MoJUIDs)	Post-clean MoJUID (unique final)	Post-clean final cum. %
1	1	Attainment	EXACT NAMES, DOB, POSTCODE, GENDER	*Exact names - MOJ Surname = NPD Surname and MOJ Forename = NPD Forename	1,388	1,260	808	0.04
1	1	HESA	EXACT NAMES, DOB, POSTCODE, GENDER	*Exact names - MOJ Surname = NPD Surname and MOJ Forename = NPD Forename	9,303	8,245	7,133	0.43
1	1	ILR	EXACT NAMES, DOB, POSTCODE, GENDER	*Exact names - MOJ Surname = NPD Surname and MOJ Forename = NPD Forename	656,915	501,533	341,245	18.95
1	1	Census	EXACT NAMES, DOB, POSTCODE, GENDER	*Exact names - 1) if NPD does not hold a middlename, MOJ Surname = NPD Surname and MOJ Forename = NPD Forename or 2) MOJ Surname = NPD Surname and MOJ Forename = NPD Forename & NPD Middlename or 3) MOJ Surname = NPD Surname and MOJ Forename = NPD Forename and MOJ MIddlenames = NPD Middlenames	1,042,296	851,768	833,655	64.20
2a	2	Census	EXACT NAMES, DOB, POSTCODE	As above without checking Gender	8,264	7,774	5,856	64.52
2a	2	ILR	EXACT NAMES, DOB, POSTCODE	As above without checking Gender	13,545	12,734	7,882	64.94
2a	2	HESA	EXACT NAMES, DOB, POSTCODE	As above without checking Gender	26,639	23,977	20,734	66.07
2a	2	Attainment	EXACT NAMES, DOB, POSTCODE	As above without checking Gender	19	17	12	66.07
2b	2	Attainment	EXACT NAMES, DOB, POSTCODE SECTOR, GENDER	As above using Postcode Sector - up to and including first character after space	59	53	40	66.07
2b	2	HESA	EXACT NAMES, DOB, POSTCODE SECTOR, GENDER	As above using Postcode Sector - up to and including first character after space	689	634	561	66.10
2b	2	Census	EXACT NAMES, DOB, POSTCODE SECTOR, GENDER	As above using Postcode Sector - up to and including first character after space	76,157	65,698	57,790	69.24
2b	2	ILR	EXACT NAMES, DOB, POSTCODE SECTOR, GENDER	As above using Postcode Sector - up to and including first character after space	65,909	52,633	36,364	71.21
2c	2	Census	STRONG FUZZY NAMES, DOB, POSTCODE, GENDER	* Strong fuzzy names - 2 versions, 1) ignoring middlenames completely MOJ Surname = NPD Surname and MOJ forename = NPD Forename and 2) MOJ Surname 'like' NPD surname or NPD Surname 'like' MOJ Surname and MOJ Forename 'like' NPD Forename or NPD Forename 'like' MOJ Forename	236,614	192,926	82,342	75.68
2c	2	ILR	STRONG FUZZY NAMES, DOB, POSTCODE, GENDER	* Strong fuzzy names - MOJ Surname 'like' NPD surname or NPD Surname 'like' MOJ Surname and MOJ Forename 'like' NPD Forename or NPD Forename 'like' MOJ Forename	89,622	66,880	40,290	77.87
2c	2	HESA	STRONG FUZZY NAMES, DOB, POSTCODE, GENDER	* Strong fuzzy names - MOJ Surname 'like' NPD surname or NPD Surname 'like' MOJ Surname and MOJ Forename 'like' NPD Forename or NPD Forename 'like' MOJ Forename	4,315	3,590	3,177	78.04
2c	2	Attainment	STRONG FUZZY NAMES, DOB, POSTCODE, GENDER	* Strong fuzzy names - MOJ Surname 'like' NPD surname or NPD Surname 'like' MOJ Surname and MOJ Forename 'like' NPD Forename or NPD Forename 'like' MOJ Forename	272	241	66	78.04
2d	2	Census	EXACT NAMES, DOB TYPO, POSTCODE, GENDER	* DOB Typo = 'month' of MOJ dob = 'month' of NPD dob and 'day' of MOJ dob = 'day' of NPD dob or 'month' of MOJ dob = 'month' of NPD dob and 'year of MOJ dob = 'year' of NPD dob or 'day' of MOJ dob = 'day' of NPD dob and 'year of MOJ dob = 'year' of NPD dob or ('month' of MOJ dob = 'day' of NPD dob and 'day' of MOJ dob = 'month' of NPD dob and 'year of MOJ dob = 'year' of NPD dob)	160,034	90,574	7,734	78.46
2d	2	HESA	EXACT NAMES, DOB TYPO, POSTCODE, GENDER	* DOB Typo = 'month' of MOJ dob = 'month' of NPD dob and 'day' of MOJ dob = 'day' of NPD dob or 'month' of MOJ dob = 'month' of NPD dob and 'year of MOJ dob = 'year' of NPD dob or 'day' of MOJ dob = 'day' of NPD dob and 'year of MOJ dob = 'year' of NPD dob or ('month' of MOJ dob = 'day' of NPD dob and 'day' of MOJ dob = 'month' of NPD dob and 'year of MOJ dob = 'year' of NPD dob)	575	412	141	78.47

Match no.	Priority	Data source	Rule	Comment	No. of matches (records)	No. of matches (MoJUIDs)	Post-clean MoJUID (unique final)	Post-clean final cum. %
2d	2	ILR	EXACT NAMES, DOB TYPO, POSTCODE, GENDER	* DOB Typo = 'month' of MOJ dob = 'month' of NPD dob and 'day' of MOJ dob = 'day' of NPD dob or 'month' of MOJ dob = 'month' of NPD dob and 'year of MOJ dob = 'year' of NPD dob or 'day' of MOJ dob = 'day' of NPD dob and 'year of MOJ dob = 'year' of NPD dob or ('month' of MOJ dob = 'day' of NPD dob and 'day' of MOJ dob = 'month' of NPD dob and 'year of MOJ dob = 'year' of NPD dob)	114,506	59,104	6,226	78.81
2d	2	Attainment	EXACT NAMES, DOB TYPO, POSTCODE, GENDER	* DOB Typo = 'month' of MOJ dob = 'month' of NPD dob and 'day' of MOJ dob = 'day' of NPD dob or 'month' of MOJ dob = 'month' of NPD dob and 'year of MOJ dob = 'year' of NPD dob or 'day' of MOJ dob = 'day' of NPD dob and 'year of MOJ dob = 'year' of NPD dob or ('month' of MOJ dob = 'day' of NPD dob and 'day' of MOJ dob = 'month' of NPD dob and 'year of MOJ dob = 'year' of NPD dob)	268	183	16	78.81
2e	2	Census	TRANSPOSED NAMES, DOB, POSTCODE, GENDER	*Transposed names - MOJ Surname = NPD Forename and MOJ Forename = NPD Surname	4,629	4,542	272	78.83
2e	2	Attainment	TRANSPOSED NAMES, DOB, POSTCODE, GENDER	*Transposed names - MOJ Surname = NPD Forename and MOJ Forename = NPD Surname	4	3	450	78.85
2e	2	HESA	TRANSPOSED NAMES, DOB, POSTCODE, GENDER	*Transposed names - MOJ Surname = NPD Forename and MOJ Forename = NPD Surname	28	28	8	78.85
2e	2	ILR	TRANSPOSED NAMES, DOB, POSTCODE, GENDER	*Transposed names - MOJ Surname = NPD Forename and MOJ Forename = NPD Surname	2,513	2,375		78.85
3a	3	Attainment	EXACT NAMES, DOB, POSTCODE (4 CHARS), GENDER	POSTCODE (4 CHARS) - first 4 characters of postcode	9	8	4	78.85
3a	3	ILR	EXACT NAMES, DOB, POSTCODE (4 CHARS), GENDER	POSTCODE (4 CHARS) - first 4 characters of postcode	31,185	25,744	19,862	79.93
3a	3	Census	EXACT NAMES, DOB, POSTCODE (4 CHARS), GENDER	POSTCODE (4 CHARS) - first 4 characters of postcode	32,324	28,503	25,799	81.33
3a	3	HESA	EXACT NAMES, DOB, POSTCODE (4 CHARS), GENDER	POSTCODE (4 CHARS) - first 4 characters of postcode	450	423	380	81.35
3b	3	HESA	EXACT NAMES, DOB, POSTCODE SECTOR	Without gender	1,390	1,273	1,124	81.41
3b	3	Attainment	EXACT NAMES, DOB, POSTCODE SECTOR	Without gender	2	1	1	81.41
3b	3	Census	EXACT NAMES, DOB, POSTCODE SECTOR	Without gender	1,250	1,221	652	81.45
3b	3	ILR	EXACT NAMES, DOB, POSTCODE SECTOR	Without gender	2,039	1,972	893	81.49
3c	3	HESA	STRONG FUZZY NAMES, DOB, POSTCODE	Without gender	9,727	8,424	7,051	81.88
3c	3	Census	STRONG FUZZY NAMES, DOB, POSTCODE	Without gender	3,800	3,644	2,117	81.99
3c	3	ILR	STRONG FUZZY NAMES, DOB, POSTCODE	Without gender	8,338	7,000	4,453	82.23
3c	3	Attainment	STRONG FUZZY NAMES, DOB, POSTCODE	Without gender	5	5	3	82.23
3d	3	HESA	EXACT NAMES, DOB TYPO, POSTCODE	Without gender	1,765	1,314	425	82.26
3d	3	Census	EXACT NAMES, DOB TYPO, POSTCODE	Without gender	408	367	104	82.26
3d	3	ILR	EXACT NAMES, DOB TYPO, POSTCODE	Without gender	2,245	1,865	490	82.29
3d	3	Attainment	EXACT NAMES, DOB TYPO, POSTCODE	Without gender	42	37	13	82.29
3e	3	Census	STRONG FUZZY NAMES, DOB, GENDER, POSTCODE SECTOR	Combo of above	22,312	19,034	10,549	82.86
3e	3	HESA	STRONG FUZZY NAMES, DOB, GENDER, POSTCODE SECTOR	Combo of above	392	347	296	82.88
3e	3	ILR	STRONG FUZZY NAMES, DOB, GENDER, POSTCODE SECTOR	Combo of above	9,534	7,306	4,476	83.12
3e	3	Attainment	STRONG FUZZY NAMES, DOB, GENDER, POSTCODE SECTOR	Combo of above	16	15		83.12
3f	3	Census	EXACT NAMES, DOB TYPO, POSTCODE SECTOR, GENDER	Combo of above	11,475	7,228	897	83.17
3f	3	HESA	EXACT NAMES, DOB TYPO, POSTCODE SECTOR, GENDER	Combo of above	141	124	53	83.17

Match no.	Priority	Data source	Rule	Comment	No. of matches (records)	No. of matches (MoJUIDs)	Post-clean MoJUID (unique final)	Post-clean final cum. %
3f	3	ILR	EXACT NAMES, DOB TYPO, POSTCODE SECTOR, GENDER	Combo of above	10,366	5,643	913	83.22
3f	3	Attainment	EXACT NAMES, DOB TYPO, POSTCODE SECTOR, GENDER	Combo of above	84	67	13	83.22
3g	3	Attainment	STRONG FUZZY NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	83	53	6	83.22
3g	3	ILR	STRONG FUZZY NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	20,250	10,130	1,030	83.28
3g	3	Census	STRONG FUZZY NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	47,929	26,499	1,726	83.37
3g	3	HESA	STRONG FUZZY NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	599	367	110	83.38
3h	3	Census	TRANSPOSED NAMES, DOB, POSTCODE	Without gender	93	92	28	83.38
3h	3	ILR	TRANSPOSED NAMES, DOB, POSTCODE	Without gender	93	92	33	83.38
3h	3	HESA	TRANSPOSED NAMES, DOB, POSTCODE	Without gender	88	88	30	83.38
3h	3	Attainment	TRANSPOSED NAMES, DOB, POSTCODE	Without gender	2	2	1	83.38
3i	3	ILR	TRANSPOSED NAMES, DOB, GENDER, POSTCODE SECTOR	Combo of above	250	235	65	83.39
3i	3	Census	TRANSPOSED NAMES, DOB, GENDER, POSTCODE SECTOR	Combo of above	270	265	22	83.39
3i	3	HESA	TRANSPOSED NAMES, DOB, GENDER, POSTCODE SECTOR	Combo of above	7	6		83.39
3j	3	ILR	TRANSPOSED NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	1,029	687	26	83.39
3j	3	Census	TRANSPOSED NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	1,875	1,254	19	83.39
3j	3	HESA	TRANSPOSED NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	4	4		83.39
3k	3	ILR	FUZZY TRANSPOSED NAMES, DOB, POSTCODE, GENDER	*Fuzzy transposed names - (MOJ Surname like NPD Forename and MOJ Forename like NPD Surname) or NPD Surname like MOJ Forename and NPD Forename like MOJ Surname)	791	711	125	83.40
3k	3	Census	FUZZY TRANSPOSED NAMES, DOB, POSTCODE, GENDER	*Fuzzy transposed names - (MOJ Surname like NPD Forename and MOJ Forename like NPD Surname) or NPD Surname like MOJ Forename and NPD Forename like MOJ Surname)	461	433	21	83.40
3k	3	HESA	FUZZY TRANSPOSED NAMES, DOB, POSTCODE, GENDER	*Fuzzy transposed names - (MOJ Surname like NPD Forename and MOJ Forename like NPD Surname) or NPD Surname like MOJ Forename and NPD Forename like MOJ Surname)	32	32	5	83.40
3k	3	Attainment	FUZZY TRANSPOSED NAMES, DOB, POSTCODE, GENDER	*Fuzzy transposed names - (MOJ Surname like NPD Forename and MOJ Forename like NPD Surname) or NPD Surname like MOJ Forename and NPD Forename like MOJ Surname)	3	3		83.40
3l	3	ILR	SURNAME, DOB, POSTCODE, GENDER	Manual check of forename	8,999	6,588		83.40
3l	3	ILR	SURNAME, PART FORENAME, DOB, POSTCODE, GENDER	Manual check of forename	42,260	31,140	6,699	83.76
3l	3	Census	SURNAME, PART FORENAME, DOB, POSTCODE, GENDER	Manual check of forename	141,858	105,548	17,363	84.71
3l	3	HESA	SURNAME, DOB, POSTCODE, GENDER	Manual check of forename	842	706	291	84.72
3l	3	Attainment	SURNAME, DOB, POSTCODE, GENDER	Manual check of forename	435	370	42	84.72
3o	3	Census	FUZZY NAMES, DOB, POSTCODE, GENDER	MOJ fullname contains both NPD Surname & NPD Forename	7,617	6,858	717	84.76
3o	3	ILR	FUZZY NAMES, DOB, POSTCODE, GENDER	MOJ fullname contains both NPD Surname & NPD Forename	3,186	2,743	264	84.78
3o	3	Attainment	FUZZY NAMES, DOB, POSTCODE, GENDER	MOJ fullname contains both NPD Surname & NPD Forename	2	2		84.78
3o	3	HESA	FUZZY NAMES, DOB, POSTCODE, GENDER	MOJ fullname contains both NPD Surname & NPD Forename	33	32	9	84.78
4a	4	Census	EXACT NAMES, DOB, POSTCODE (3 CHARS), GENDER	POSTCODE (3CHARS) - first 3 characters of postcode	14,806	13,044	11,834	85.42
4a	4	ILR	EXACT NAMES, DOB, POSTCODE (3 CHARS), GENDER	POSTCODE (3CHARS) - first 3 characters of postcode	13,725	11,381	8,748	85.89
4a	4	HESA	EXACT NAMES, DOB, POSTCODE (3 CHARS), GENDER	POSTCODE (3CHARS) - first 3 characters of postcode	271	252	233	85.91

Match no.	Priority	Data source	Rule	Comment	No. of matches (records)	No. of matches (MoJUIDs)	Post-clean MoJUID (unique final)	Post-clean final cum. %
4a	4	Attainment	EXACT NAMES, DOB, POSTCODE (3 CHARS), GENDER	POSTCODE (3CHARS) - first 3 characters of postcode	4	4	3	85.91
4b	4	ILR	EXACT NAMES, DOB, POSTCODE (4 CHARS)	Combo of above	1,383	1,346	607	85.94
4b	4	HESA	EXACT NAMES, DOB, POSTCODE (4 CHARS)	Combo of above	1,035	943	841	85.99
4b	4	Census	EXACT NAMES, DOB, POSTCODE (4 CHARS)	Combo of above	760	749	339	86.00
4c	4	Attainment	STRONG FUZZY NAMES, DOB, GENDER, POSTCODE (4 CHARS)	Combo of above	6	5	2	86.00
4c	4	HESA	STRONG FUZZY NAMES, DOB, GENDER, POSTCODE (4 CHARS)	Combo of above	207	180	156	86.01
4c	4	Census	STRONG FUZZY NAMES, DOB, GENDER, POSTCODE (4 CHARS)	Combo of above	14,073	12,636	6,926	86.39
4c	4	ILR	STRONG FUZZY NAMES, DOB, GENDER, POSTCODE (4 CHARS)	Combo of above	5,914	4,758	2,966	86.55
4d	4	HESA	EXACT NAMES, DOB TYPO, POSTCODE(4 CHARS), GENDER	Combo of above	39	34	12	86.55
4d	4	ILR	EXACT NAMES, DOB TYPO, POSTCODE(4 CHARS), GENDER	Combo of above	6,193	3,432	638	86.58
4d	4	Census	EXACT NAMES, DOB TYPO, POSTCODE(4 CHARS), GENDER	Combo of above	2,887	2,174	439	86.61
4d	4	Attainment	EXACT NAMES, DOB TYPO, POSTCODE(4 CHARS), GENDER	Combo of above	24	14	3	86.61
4e	4	ILR	FUZZY TRANSPOSED NAMES, DOB, POSTCODE	Without Gender	30	30	14	86.61
4e	4	Census	FUZZY TRANSPOSED NAMES, DOB, POSTCODE	Without Gender	10	10	3	86.61
4e	4	HESA	FUZZY TRANSPOSED NAMES, DOB, POSTCODE	Without Gender	74	71	26	86.61
4f	4	Census	FUZZY TRANSPOSED NAMES, DOB, GENDER, POSTCODE SECTOR	Combo of above	44	39	2	86.61
4f	4	HESA	FUZZY TRANSPOSED NAMES, DOB, GENDER, POSTCODE SECTOR	Combo of above	2	2	1	86.61
4f	4	ILR	FUZZY TRANSPOSED NAMES, DOB, GENDER, POSTCODE SECTOR	Combo of above	84	79	19	86.61
4g	4	Attainment	FUZZY TRANSPOSED NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	1	1	1	86.61
4g	4	Census	FUZZY TRANSPOSED NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	176	115	2	86.61
4g	4	ILR	FUZZY TRANSPOSED NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	268	193	15	86.61
4g	4	HESA	FUZZY TRANSPOSED NAMES, DOB TYPO, POSTCODE, GENDER	Combo of above	8	6		86.61
5a	5	Census	EXACT NAMES, DOB, POSTCODE (3 CHARS)	Without Gender	7,482	6,983	4,189	86.84
5a	5	Attainment	EXACT NAMES, DOB, POSTCODE (3 CHARS)	Without Gender	1	1		86.84
5a	5	ILR	EXACT NAMES, DOB, POSTCODE (3 CHARS)	Without Gender	369	363	159	86.85
5a	5	HESA	EXACT NAMES, DOB, POSTCODE (3 CHARS)	Without Gender	613	544	490	86.88
5b	5	Census	STRONG FUZZY NAMES, DOB, GENDER, POSTCODE (3 CHARS)	Combo of above	1,567	1,357	555	86.91
6a	6	Census	EXACT NAMES, DOB, POSTCODE (2 CHARS), GENDER	POSTCODE (2CHARS) - first 2 characters of postcode - as this postcode check is quite fuzzy, also check if only one instance of a pupilid with that combination of names and DOB	31,486	27,780	25,286	88.28
6a	6	Attainment	EXACT NAMES, DOB, POSTCODE (2 CHARS), GENDER	POSTCODE (2CHARS) - first 2 characters of postcode - as this postcode check is quite fuzzy, also check if only one instance of a pupilid with that combination of names and DOB	17	15	11	88.28
6a	6	ILR	EXACT NAMES, DOB, POSTCODE (2 CHARS), GENDER	POSTCODE (2CHARS) - first 2 characters of postcode - as this postcode check is quite fuzzy, also check if only one instance of a pupilid with that combination of names and DOB	27,466	23,282	18,185	89.27
6a	6	HESA	EXACT NAMES, DOB, POSTCODE (2 CHARS), GENDER	POSTCODE (2CHARS) - first 2 characters of postcode - as this postcode check is quite fuzzy, also check if only one instance of a pupilid with that combination of names and DOB	631	591	539	89.30

Match no.	Priority	Data source	Rule	Comment	No. of matches (records)	No. of matches (MoJUIDs)	Post-clean MoJUID (unique final)	Post-clean final cum. %
7a	7	Attainment	FORENAME, SURNAME, DOB	Exact match on Forename, Surname & DOB against a record where only one instance of a pupilid with that combination of names and DOB.	130,654	110,517	94,720	94.44
7a	7	ILR	FORENAME, SURNAME, DOB	Exact match on Forename, Surname & DOB against a record where only one instance of a pupilid with that combination of names and DOB.	45,966	36,812	28,384	95.98
7a	7	Census	FORENAME, SURNAME, DOB	Exact match on Forename, Surname & DOB against a record where only one instance of a pupilid with that combination of names and DOB.	92,660	80,101	69,520	99.75
7a	7	HESA	FORENAME, SURNAME, DOB	Exact match on Forename, Surname & DOB against a record where only one instance of a pupilid with that combination of names and DOB.	5,761	5,287	4,608	100.00
					3,349,650	2,608,820	1,842,478	

In the table above, the sixth column 'number of matches (records)' shows the number of matches, including one-to-many, many-to-one, many-to-many and one-to-one. Each match is counted separately, for example, an MoJUID matched to 7 education IDs (PMRs) is counted as 7. The table below shows the MoJUIDs as unique.

	All UIDs		PNC UIDs only	
PMR count	**Number of UIDs**	**Percent**	**Number of UIDs**	**Percent**
1	1,237,186	64.5%	887,904	57.2%
2	410,200	21.4%	395,130	25.5%
3	113,216	5.9%	112,384	7.2%
4	72,877	3.8%	72,781	4.7%
5	22,285	1.2%	22,275	1.4%
6	23,837	1.2%	23,837	1.5%
7 or more	38,224	2.0%	38,224	2.5%
Total	**1,917,825**	100.0%	**1,552,535**	100.0%

These counts above are the MoJUID counts used by MoJ to report on cleaning steps using the source dataset to select the match.

Appendix D: Additional results

Table D1: School census (pupil-level) data

Cohort born	Data available for school year:											
	R	1	2	3	4	5	6	7	8	9	10	11
1985/86												X
1986/87											X	X
1987/88										X	X	X
1988/89									X	X	X	X
1989/90								X	X	X	X	X
1990/91							X	X	X	X	X	X
1991/92						X	X	X	X	X	X	X
1992/93					X	X	X	X	X	X	X	X
1993/94				X	X	X	X	X	X	X	X	X
1994/95			X	X	X	X	X	X	X	X	X	X
1995/96		X	X	X	X	X	X	X	X	X	X	X
1996/97	X	X	X	X	X	X	X	X	X	X	X	X
1997/98	X	X	X	X	X	X	X	X	X	X	X	X
1998/99	X	X	X	X	X	X	X	X	X	X	X	X
1999/00	X	X	X	X	X	X	X	X	X	X	X	X
2000/01	X	X	X	X	X	X	X	X	X	X	X	X
2001/02	X	X	X	X	X	X	X	X	X	X	X	X
2002/03	X	X	X	X	X	X	X	X	X	X	X	X
2003/04	X	X	X	X	X	X	X	X	X	X	X	
2004/05	X	X	X	X	X	X	X	X	X	X		
2005/06	X	X	X	X	X	X	X	X	X			
2006/07	X	X	X	X	X	X	X	X				

Table D2: Attainment data: early years to Key Stage 4

Cohort born	Early years	KS1	KS2	KS3	KS4
1985/86			X	X	X
1986/87			X	X	X
1987/88			X	X	X
1988/89			X	X	X
1989/90			X	X	X
1990/91		X	X	X	X
1991/92		X	X	X	X
1992/93		X	X	X	X
1993/94		X	X	X	X
1994/95		X	X	X	X
1995/96		X	X	X	X
1996/97		X	X	X	X
1997/98	X	X	X	X	X
1998/99	X	X	X	X	X
1999/00	X	X	X		X
2000/01	X	X	X		X
2001/02	X	X	X		X
2002/03	X	X	X		X
2003/04	X	X	X		
2004/05	X	X	X		
2005/06	X	X	X		
2006/07	X	X	X		

Absence data are available for academic year 2005/2006 onwards. Thus, for those born between 1985/86 and 1988/89 there is no absence data. This is summarised in Table D3.

Table D3: Absence data

Cohort born	R	1	2	3	4	5	6	7	8	9	10	11
1985/86												
1986/87												
1987/88												
1988/89												
1989/90												X
1990/91											X	X
1991/92										X	X	X
1992/93									X	X	X	X
1993/94								X	X	X	X	X
1994/95							X	X	X	X	X	X
1995/96						X	X	X	X	X	X	X
1996/97					X	X	X	X	X	X	X	X
1997/98				X	X	X	X	X	X	X	X	X
1998/99			X	X	X	X	X	X	X	X	X	X
1999/00		X	X	X	X	X	X	X	X	X	X	X
2000/01	X	X	X	X	X	X	X	X	X	X	X	X
2001/02	X	X	X	X	X	X	X	X	X	X	X	X
2002/03	X	X	X	X	X	X	X	X	X	X	X	X
2003/04	X	X	X	X	X	X	X	X	X	X	X	
2004/05	X	X	X	X	X	X	X	X	X	X		
2005/06	X	X	X	X	X	X	X	X	X			
2006/07	X	X	X	X	X	X	X	X				

In Table D4, the years highlighted in yellow are those containing the limited set of exclusion data (2001/02-2004/05), as outlined in Section 2 of the report; the remainder contain the full set of data.

Table D4: Exclusions data

Cohort born	R	1	2	3	4	5	6	7	8	9	10	11
1985/86												X
1986/87											X	X
1987/88										X	X	X
1988/89									X	X	X	X
1989/90								X	X	X	X	X
1990/91							X	X	X	X	X	X
1991/92						X	X	X	X	X	X	X
1992/93					X	X	X	X	X	X	X	X
1993/94				X	X	X	X	X	X	X	X	X
1994/95			X	X	X	X	X	X	X	X	X	X
1995/96		X	X	X	X	X	X	X	X	X	X	X
1996/97	X	X	X	X	X	X	X	X	X	X	X	X
1997/98	X	X	X	X	X	X	X	X	X	X	X	X
1998/99	X	X	X	X	X	X	X	X	X	X	X	X
1999/00	X	X	X	X	X	X	X	X	X	X	X	X
2000/01	X	X	X	X	X	X	X	X	X	X	X	X
2001/02	X	X	X	X	X	X	X	X	X	X	X	X
2002/03	X	X	X	X	X	X	X	X	X	X	X	X
2003/04	X	X	X	X	X	X	X	X	X	X	X	
2004/05	X	X	X	X	X	X	X	X	X	X		
2005/06	X	X	X	X	X	X	X	X	X			
2006/07	X	X	X	X	X	X	X	X				

Table D5: Looked After Children (CLA) data

Cohort born	R	1	2	3	4	5	6	7	8	9	10	11
1985/86												
1986/87												
1987/88												
1988/89												
1989/90												X
1990/91											X	X
1991/92										X	X	X
1992/93									X	X	X	X
1993/94								X	X	X	X	X
1994/95							X	X	X	X	X	X
1995/96						X	X	X	X	X	X	X
1996/97					X	X	X	X	X	X	X	X
1997/98				X	X	X	X	X	X	X	X	X
1998/99			X	X	X	X	X	X	X	X	X	X
1999/00		X	X	X	X	X	X	X	X	X	X	X
2000/01	X	X	X	X	X	X	X	X	X	X	X	X
2001/02	X	X	X	X	X	X	X	X	X	X	X	X
2002/03	X	X	X	X	X	X	X	X	X	X	X	
2003/04	X	X	X	X	X	X	X	X	X	X		
2004/05	X	X	X	X	X	X	X	X	X			
2005/06	X	X	X	X	X	X	X	X				
2006/07	X	X	X	X	X	X	X					

Table D6: Children in Need (CiN) data

Cohort born	R	1	2	3	4	5	6	7	8	9	10	11
1985/86												
1986/87												
1987/88												
1988/89												
1989/90												
1990/91												
1991/92												
1992/93												X
1993/94											X	X
1994/95										X	X	X
1995/96									X	X	X	X
1996/97								X	X	X	X	X
1997/98							X	X	X	X	X	X
1998/99						X	X	X	X	X	X	X
1999/00					X	X	X	X	X	X	X	X
2000/01				X	X	X	X	X	X	X	X	X
2001/02			X	X	X	X	X	X	X	X	X	X
2002/03		X	X	X	X	X	X	X	X	X	X	X
2003/04	X	X	X	X	X	X	X	X	X	X	X	
2004/05	X	X	X	X	X	X	X	X	X	X		
2005/06	X	X	X	X	X	X	X	X	X			
2006/07	X	X	X	X	X	X	X	X				

Table D7: Number of pupils in each birth year and percentage of these in the pupil level census for each year (using school census data up to 2017/18)

Cohort born	Total number of individuals	R	1	2	3	4	5	6	7	8	9	10	11
1985/86	653,515	X	X	X	X	X	X	X	X	X	X	X	82%
1986/87	673,884	X	X	X	X	X	X	X	X	X	X	84%	85%
1987/88	691,920	X	X	X	X	X	X	X	X	X	84%	87%	85%
1988/89	685,942	X	X	X	X	X	X	X	X	84%	87%	86%	85%
1989/90	696,672	X	X	X	X	X	X	X	86%	87%	87%	87%	85%
1990/91	719,977	X	X	X	X	X	X	83%	85%	85%	85%	85%	85%
1991/92	716,146	X	X	X	X	X	82%	86%	85%	85%	85%	86%	85%
1992/93	692,280	X	X	X	X	82%	86%	85%	85%	85%	86%	86%	85%
1993/94	690,696	X	X	X	82%	86%	85%	86%	85%	86%	86%	86%	86%
1994/95	677,937	X	X	81%	85%	85%	85%	86%	85%	86%	86%	86%	86%
1995/96	675,177	X	81%	84%	84%	85%	85%	85%	85%	85%	85%	86%	85%
1996/97	691,482	79%	84%	84%	84%	84%	85%	85%	84%	84%	85%	85%	85%
1997/98	686,336	80%	82%	82%	83%	83%	83%	83%	83%	83%	83%	84%	83%
1998/99	683,065	79%	82%	82%	83%	83%	83%	83%	82%	83%	83%	83%	83%
1999/00	668,269	78%	82%	82%	82%	82%	83%	83%	82%	82%	83%	83%	83%
2000/01	652,264	80%	83%	83%	83%	83%	83%	83%	83%	83%	84%	84%	84%
2001/02	639,685	83%	84%	84%	84%	84%	84%	84%	84%	84%	85%	85%	84%
2002/03	656,386	84%	85%	85%	85%	85%	85%	85%	85%	85%	86%	85%	X
2003/04	675,224	84%	85%	85%	85%	85%	85%	86%	85%	86%	85%	X	X
2004/05	688,513	84%	85%	85%	85%	85%	86%	86%	86%	86%	X	X	X
2005/06	697,826	85%	86%	86%	86%	86%	87%	87%	86%	X	X	X	X
2006/07	716,095	86%	87%	87%	87%	87%	88%	87%	X	X	X	X	X

Table D8: Completeness of NC year, FSM, SEN status, and IDACI scores in the main school census

Academic year	Term	Number of records	Number of unique PMR	Number with missing or invalid			
				NC year	FSM	SEN	IDACI
2001/02	-	7,097,264	7,094,211	201	5,878	1,733	85,846 (1.2%)
2002/03	-	7,283,265	7,280,817	93	0	12,041	42,701 (0.6%)
2003/04	-	7,399,237	7,397,460	0	0	9,890	35,297 (0.5%)
2004/05	-	7,363,232	7,361,832	131	0	659	31,809 (0.4%)
2005/06	Spring	7,308,272	7,307,407	0	0	0	33,891 (0.5%)
	Summer	3,328,855	3,328,299	0	0	0	2,362,574 (71.0%)
2006/07	Spring	7,241,771	7,241,013	2,200	0	0	34,391 (0.5%)
	Summer	7,228,738	7,227,350	2,282	0	0	33,743 (0.5%)
	Autumn	3,338,221	3,337,107	38	0	300	NR[2]
2007/08	Spring	7,187,779	7,187,183	1,867	0	0	31,230 (0.4%)
	Summer	7,172,610	7,171,881	1,799	0	0	30,687 (0.4%)
	Autumn	7,143,557	7,140,823	<10	0	0	35,881 (0.5%)
2008/09	Spring	7,155,606	7,155,399	0	0	0	27,959 (0.4%)
	Summer	7,146,958	7,146,367	0	0	0	27,372 (0.4%)
	Autumn	7,105,708	7,104,368	<10	0	0	31,666 (0.4%)
2009/10	Spring	7,170,084	7,169,848	0	0	0	24,409 (0.3%)
	Summer	7,160,646	7,160,191	0	0	0	23,677 (0.3%)
	Autumn	7,128,647	7,127,578	0	0	0	27,834 (0.4%)
2010/11	Spring	7,191,189	7,190,988	0	0	NR[2]	22,179 (0.3%)
	Summer	7,178,493	7,178,223	<10	0	0	NR[2]
	Autumn	7,152,025	7,151,287	0	0	0	25,604 (0.4%)
2011/12	Spring	7,237,061	7,236,771	0	0	0	20,511 (0.3%)
	Summer	7,223,671	7,223,301	0	0	0	25,604 (0.4%)
	Autumn	7,227,807	7,227,375	0	0	<10	NR[2]
2012/13[1]	Spring	6,678,111	6,677,802	0	0	0	18,125 (0.3%)
	Summer	6,668,031	6,667,667	0	0	0	18,499 (0.3%)
	Autumn	6,685,141	6,684,621	0	0	0	24,657 (0.4%)
2013/14	Spring	6,129,494	6,128,898	0	0	<10	15,040 (0.2%)
	Summer	6,121,547	6,120,726	0	0	0	16,263 (0.3%)
	Autumn	6,122,918	6,122,436	0	0	0	19,672 (0.3%)
2014/15	Spring	5,574,148	5,573,550	0	0	0	13,226 (0.2%)
	Summer	5,566,680	5,565,872	0	0	0	14,428 (0.3%)
	Autumn	5,577,939	5,577,197	0	0	<10	17,360 (0.3%)
2015/16	Spring	5,010,781	5,010,237	0	0	0	11,021 (0.2%)
	Summer	5,003,206	5,002,516	0	0	0	11,734 (0.2%)
	Autumn	5,014,225	5,013,440	0	0	<10	15,786 (0.3%)
2016/17	Spring	4,451,188	4,450,767	<10	0	0	8,801 (0.2%)
	Summer	4,440,374	4,439,786	<10	0	<10	9,281 (0.2%)
	Autumn	4,458,606	4,457,885	0	0	47	12,117 (0.3%)
2017/18	Spring	3,884,176	3,883,631	<10	0	0	6,954 (0.2%)
	Summer	3,873,836	3,873,298	0	0	0	7,380 (0.2%)
	Autumn	3,893,513	3,892,860	0	0	0	9,836 (0.3%)

1. The latest reception year for our cohort (those born September 1985 – August 2007) is 2011/2012; therefore, from 2012/2013, the sample size in the census data begins to decline
2. NR = not recorded

Table D9: Completeness of NC year, FSM indicator, SEN status and IDACI score in the AP census data

Academic years	Number of records	Number of unique PMR	Number with missing or invalid			
			NC year	FSM	SEN	IDACI
2007/08 – 2019/20	284,340	116,992	0	921 (0.8%)	17 (<0.1%)	Not recorded

Table D10: Completeness of NC year, FSM indicator, SEN status and IDACI score in the PRU census data

Academic years	Number of records	Number of unique PMR	Number with missing or invalid			
			NC year	FSM	SEN	IDACI
2009/10 – 2012/13	56,015	42,467	0	0	<10	640 (1.1%)

Table D11: Completeness in absence data: percentage with calculable value

Cohort born	1	2	3	4	5	6	7	8	9	10	11
1989/90	X	X	X	X	X	X	X	X	X	X	97.5%
1990/91	X	X	X	X	X	X	X	X	X	98.0%	98.0%
1991/92	X	X	X	X	X	X	X	X	98.1%	99.0%	97.5%
1992/93	X	X	X	X	X	X	X	98.3%	99.1%	98.8%	97.6%
1993/94	X	X	X	X	X	X	96.9%	99.3%	99.0%	98.9%	96.8%
1994/95	X	X	X	X	X	X	99.4%	99.1%	99.1%	98.6%	96.8%
1995/96	X	X	X	X	X	99.8%	99.2%	99.2%	99.0%	98.6%	97.5%
1996/97	X	X	X	X	99.8%	99.5%	99.3%	99.2%	98.9%	99.8%	97.4%
1997/98	X	X	X	99.8%	99.4%	99.5%	99.3%	99.2%	99.3%	98.8%	98.4%
1998/99	X	X	99.8%	99.4%	99.5%	99.5%	99.3%	99.5%	99.2%	99.2%	98.5%
1999/00	X	99.8%	99.4%	99.5%	99.5%	99.5%	99.6%	99.5%	99.4%	99.3%	98.1%
2000/01	99.8%	99.4%	99.5%	99.5%	99.5%	99.8%	99.6%	99.5%	99.4%	99.2%	98.5%
2001/02	99.4%	99.5%	99.5%	99.5%	99.8%	99.8%	99.6%	99.6%	99.4%	99.3%	99.0%
2002/03	99.5%	99.5%	99.5%	99.8%	99.8%	99.8%	99.6%	99.5%	99.4%	99.3%	X
2003/04	99.4%	99.5%	99.9%	99.8%	99.8%	99.8%	99.6%	99.6%	99.4%	X	X
2004/05	99.3%	99.7%	99.7%	99.7%	99.6%	99.5%	99.5%	99.4%	X	X	X
2005/06	99.7%	99.7%	99.7%	99.7%	99.7%	99.6%	99.5%	X	X	X	X
2006/07	99.7%	99.7%	99.7%	99.7%	99.7%	99.7%	X	X	X	X	X

Table D12: Number appearing more than once (in different academic years) in the school census for a particular school year

Cohort born	R	1	2	3	4	5	6	7	8	9	10	11
1985/86	X	X	X	X	X	X	X	X	X	X	X	1,239
1986/87	X	X	X	X	X	X	X	X	X	X	3,563	5,745
1987/88	X	X	X	X	X	X	X	X	X	3,286	7,865	982
1988/89	X	X	X	X	X	X	X	X	3,166	6,849	1,044	820
1989/90	X	X	X	X	X	X	X	3,371	6,591	789	1,110	1,189
1990/91	X	X	X	X	X	X	1,042	4,660	982	825	1,646	870
1991/92	X	X	X	X	X	3,812	8,651	1,154	2,673	1,575	1,239	1,299
1992/93	X	X	X	X	3,572	9,314	1,348	636	1,276	707	1,045	1,236
1993/94	X	X	X	4,159	8,817	1,433	1,480	507	753	668	1,254	1,134
1994/95	X	X	3,857	9,781	1,670	1,181	1,698	519	487	715	1,075	850
1995/96	X	4,913	9,855	1,724	1,660	1,711	674	352	527	593	1,191	937
1996/97	4,549	9,583	2,108	1,353	1,970	679	707	442	459	630	1,365	1,019
1997/98	7,305	2,638	1,851	1,710	609	560	602	324	460	642	1,607	991
1998/99	3,334	4,427	3,168	684	595	568	577	297	445	931	1,503	914
1999/00	3,574	7,825	1,222	550	520	426	537	331	721	759	1,338	714
2000/01	3,367	2,564	830	516	394	383	422	547	439	565	1,420	618
2001/02	2,832	850	657	413	353	378	509	280	446	610	1,171	113
2002/03	4,066	817	589	365	347	455	349	330	436	435	272	X
2003/04	5,188	848	529	334	438	482	429	324	397	94	X	X
2004/05	3,114	911	452	487	441	451	374	265	88	X	X	X
2005/06	2,824	1,010	609	548	522	495	384	57	X	X	X	X
2006/07	2,280	906	465	508	526	400	124	X	X	X	X	X

Table D13: Numbers appearing more than once in a given school census with two different national curriculum years

Year	Term	Number	Term	Season	Number
2001/02	-	400	2011/12	Spring	18
2002/03	-	159		Summer	23
2003/04	-	130		Autumn	59
2004/05	-	143	2012/13	Spring	12
2005/06	Spring	58		Summer	20
	Summer	34		Autumn	38
2006/07	Spring	57	2013/14	Spring	41
	Summer	94		Summer	69
	Autumn	131		Autumn	54
2007/08	Spring	45	2014/15	Spring	51
	Summer	52		Summer	53
	Autumn	124		Autumn	79
2008/09	Spring	15	2015/16	Spring	26
	Summer	44		Summer	29
	Autumn	102		Autumn	64
2009/10	Spring	10	2016/17	Spring	19
	Summer	20		Summer	22
	Autumn	70		Autumn	66
2010/11	Spring	<10	2017/18	Spring	29
	Summer	14		Summer	33
	Autumn	43		Autumn	60

Acknowledgements

This project was commissioned by ADR UK (Administrative Data Research UK) on behalf of the Home Office, to assess the feasibility of using the MoJ-DfE linked dataset for evaluating early interventions for violence prevention. It is funded by ADR UK. ADR UK is funded by the Economic and Social Research Council (part of UK Research and Innovation).

Authors

Amy Dillon[1], Andy Boyd[1], Mark Mummé[1], Kate Tilling[1, 2], Iain Brennan[3], and Rosie Cornish[1, 2]

1. Population Health Sciences, Bristol Medical School, University of Bristol

2. MRC Integrative Epidemiology Unit, University of Bristol

3. Department of Criminology and Sociology, University of Hull

Visit the ADR UK website

@ADR_UK