LSYPE2 User Guide to the Datasets: wave 1

September 2015

TNS BMRB

Contents

1. Introduction 3

2. Survey Content 4

2.1. Wave 1 deposited data 4

2.2. How to link the datasets 4

2.3. Multicoded variables 5

2.4. Missing values 5

2.5. Variable names 8

2.6. Variable and value labels 9

2.7. Data cleaning 9

2.8. Datasets 10

3. Specifying the sample design 12

4. Linked Data 13

5. Questionnaire and Data Problems 18

1. Introduction

This user guide provides detailed information pertaining to data arising from wave 1 of the second cohort of the Longitudinal Study of Young People in England (LSYPE2), managed by the Department for Education (DfE). These data are available to download from the UK Data Service.

This user guide only contains information about the datasets themselves. For further information about the background to the survey, the fieldwork procedures, sampling, and weighting please refer to the technical report which is also available to download from the UK Data Service.

This user guide was created jointly by TNS BMRB and DfE.

2. Survey Content

2.1. Wave 1 deposited data

The LSYPE2 Wave 1 datsets were deposited in September 2015.

There are three levels of dataset deposited on the Data Service:

1. Safeguarded data – containing the majority of the variables from the main parent, second parent and young person interviews as well as many of the derived variables and some high-level geodemographic and sampling data. This contains anonymised versions of variables (discussed in section 2.8).

2. Controlled access (remote access) – for the survey data, this has similar content to the safeguarded dataset, additionally containing the majority of sensitive derived, sample, geodemographic and survey variables which are excluded from the safeguarded file. Some of the most sensitive variables remain anonymised in this file. This is accompanied by three files of NPD data, which exclude all sensitive variables:

    - School-level census data about the school the young person attended, from 2006, 2010 and 2013 (i.e. the years they completed KS1 and KS2, plus wave 1). This also includes Ofsted ratings and geodemographic data.
    - Pupil-level data about the young person's KS1 attainment, from 2006.
    - School-level data about the KS1 and KS4 levels of attainment in the school the young person attended, from 2006 and 2013 (W1 school) respectively.

    In addition there are four datasets containing the most sensitive survey and geodemographic variables, covering:

    - Detailed characteristics
    - Income
    - Health
    - Care

3. Controlled access (available through the safe room only)– this contains one pupil-level NPD file, containing particularly sensitive information about the young person such as their ethnicity.

The various NPD files are discussed further in section 4, as is what is meant by safeguarded and controlled data.

Finally, DfE hold one survey file that includes a primary sampling unit code for all individuals, together with the survey data only. This file will only be made available upon application to the DfE. Please contact team.longitudnal@education.gsi.gov.uk for more information.

2.2. How to link the datasets

All of the datasets have a unique serial number and each file can be linked on the variable - surveyID_W1_ADM. This serial number is unique to the cohort member and therefore each household. It is important that each file is sorted by this surveyID in ascending order to link the datasets. A typical SPSS command to link files is shown below.

Merging datasets example:

```
GET FILE='C:\ LSYPE2 Wave 1_safeguarded release v7.sav'
Sort cases by surveyID_W1_ADM (A).
SAVE OUTFILE='C:\ LSYPE2 Wave 1_safeguarded release v7.sav'
GET FILE='C:\ LSYPE2 Wave 1_health_controlled_remoteaccess.sav'
Sort cases by surveyID_W1_ADM (A).
SAVE OUTFILE='C:\ LSYPE2 Wave 1_health_controlled_remoteaccess v2.sav'

GET FILE='C:\ LSYPE2 Wave 1_safeguarded_release v7.sav'
MATCH FILES /FILE=*
/FILE='C:\ LSYPE2 Wave 1_health_controlled_remoteaccess v2.sav'
/BY surveyID_W1_ADM
SAVE OUTFILE='C:\ LSYPE2 Wave 1_safeguarded and controlledhealth.sav'.
EXECUTE.
```

2.3. Multicoded variables

Multicoded variables are obtained from questions where the interviewer is instructed to 'code all that apply'. Each response category has a separate variable in the dataset. For example, benefits received by the main parent (and their partner) have been stored within the datasets as multicoded variables, therefore if a main parent has answered that they receive Jobseekers Allowance and Child Benefit then they will have a 'yes' response in both of these separate variables.

2.4. Missing values

Due to the complexity of the information collected during the survey, a number of missing value categories have been adopted. These are shown in Box 1. (Derived variable codes are covered by the separate derived variable documentation.)

Box 1: Summary of missing values applied to the LSYPE2 survey and sampling data.

Valid Missing Values (generally included within calculated percentages)

-1 Don't know – enables respondents to answer don't know to questions.

-92 Refused – used to signify when a respondent has refused to answer a particular question.

-2 Don't know / refused – used in some edited variables to represent a response of either of the above codes

Invalid Missing Values (generally excluded from calculated percentages)

-3 Data unavailable – used in some edited variables to represent data that is not available

-4 Independent/PRU – used in sampling variables to denote pupils sampled from independent schools or pupil referral units, for whom sampling characteristics data is not available

-91 Not applicable – used to signify that a question did not apply to a respondent, usually due to routing.

-93 Interview terminated early – used to signify cases where the respondent got a significant way through the interview but did not complete to the end

-94 Not enough information – used for SOC coding and signifies that there is not enough information for the response to be classfied

-95 Telephone recontact – used to signify cases for which some data was recollected through a telephone recontact exercise but for which a particular question was not relevant or appropriate to ask

-96 History respondent misidentified – used to signify that a history respondent was present but the wrong person in the household was identified as the history respondent and completed the interview due to a technical error

-97 Data missing – used to signify cases where the respondent should have answered a question but didn't (or the data has been lost) due to either a technical error, a script error or a coding error during the interview

-98 Not present - used to signify that a respondent was not identified for this part of the questionnaire module (i.e. respondent was a single parent).

-99 Respondent not interviewed - used to signify that a respondent was identified as eligible to answer the relevant questionnaire modules but was not interviewed (this may be due to a number of reasons, i.e. not being available or refusing to take part).

In the NPD and geodemographic variables, missing data (for whatever reason) is primarily represented by -987, though there are a small number of specific missing or unsubstantive value codes in some variables:

Box 2: Summary of missing values in NPD data

Across all NPD and geodemographic variables

987 = Data missing. [This can cover a variety of reasons in a single variable.]

In some school KS4 cohort attainment variables

-995 or NA = Not applicable or not available.

-994 or NE = Not entered [in the context of an exam].

-991 or NP = Not published.

-990 = Low coverage [<50% of pupils included in calculation].

-989 or SUPP = Suppressed at publication [due to small numbers].

In some KS1 pupil attainment variables

A = Absent.

D = Disapplied from the National Curriculum.

U = Unable to access.

IN = Invalid.

In some pupil census variables

INVA or INV = Invalid value

MISS or MIS = Missing value

N/A = Not available

NOBT or NOT = Not yet obtained

REFU or REF = Refused

UNCL = Unclassified

PEN = Classification pending

2.5. Variable names

All survey variable names follow the name given to the variable in the word questionnaire and have been given up to four suffixes. This is to:

- enable users to clearly distinguish between the different waves of data for both cross-sectional and longitudinal analysis
- enable users to clearly distinguish between the different modules of the interview completed by the young person, the main parent or the second parent
- identify edited variables and distinguish between versions in different levels of dataset

Each survey variable name in the data has been named using the variable name which directly relates to the questionnaire, followed by a numeric suffix i.e. _1, _2 for multi-coded variables, a suffix to identify the wave of the survey and then a suffix to identify which questionnaire module the variable comes from, i.e. the household grid, the main parent module, the second parent module, the history module or the young person interview. (Sampling, geodemographic and derived variables follow the same structure, with question name replaced by a description of the content. Derived variables relating to mothers or fathers are denoted by adding M or F respectively directly on to the end of question name, and versions for MPs and SPs are denoted by adding either MP/SP or _MP/_SP respectively.)

Derived variables, administrative variables and sampling variables have been given different suffixes to identify them.

Unedited multicoded variable names are made up of the following characters:

[Question name] [Suffix1] [Suffix2] [Suffix3]

A typical unedited single coded variable name is made up of the following characters:

[Question name] [Suffix2] [Suffix3]

Question name	is directly comparable with the questionnaire. It is easy to search for questions within your dataset.
Suffix1	is numerical to indicate a multicoded variable, starting from 1 for the first answer and continuing sequentially
Suffix2	indicates the wave - W1= wave 1; W2 = wave 2 etc.
Suffix 3	indicates which module of the questionnaire the question was asked in: GRID = household grid; MP = the Main Parent module or Main Parent questions in the Individual Parent module; SP = the Second Parent questions in the Individual Parent module; HIST = the History module; YP = the Young Person interview. For variables not directly asked in the interview Suffix 3 will be as follows; DER = derived variable[1]; ADM = administrative variable; SAM = sample variable; GEO = geodemographic variable (based on sample address).

[1] Full details of all derived variables are available in the 'LSYPE2 Wave 1 Derived Variable Documentation' which has also been deposited.

Edited variables are denoted by the inclusion of a suffix of _A directly before suffix 2; these have undergone one or more edits to protect respondents, because of small numbers or sensitive data.

NPD variable names (and those derived from NPD data) have been minimally changed from the original source variables. In general, the name will contain a short description of the content, plus an indication of the year described (e.g. LEA06_age_6_pupils contains the number of age 6 pupils in the school in 2006).

Variable names may mirror those from the first LSYPE, data from which is also available through the UK Data Service. This does not necessarily mean that the variable construction is identical, only that the questions in each survey had the same names: for example, Pladk16 had its codeset updated between waves to reflect changes in post-16 options, and Cignow changed question wording. To assess the comparability of variables between LSYPE cohorts, please consult the published questionnaires.

2.6. Variable and value labels

The survey variable labels included on the datasets are intended to be a full description of what the variable is asking about but do not include the exact wording used in the questionnaire or a description of which groups were asked a particular question.[2] Similarly, the derived, sampling and geodemographic variable labels are intended to describe the content of the variable but not the coverage. The NPD variable labels have mostly been taken directly from the NPD documentation, although edits have been made to explicitly state the cohort described and clarify the most obscure.

Value labels have been appended for all variables, to show what each value represents. If no value label is appended for a response (and an error is not noted in the accompanying spreadsheet – see section 5), it should be interpreted literally.

2.7. Data cleaning

The survey data has gone through an extensive process of checks to ensure the consistency and validity of the data. These are checks that investigate any outliers found within the data, ensure that the data has followed the routing used in the questionnaire, ensure that the correct person has answered the relevant questions and ensure that information is consistent between directly comparable variables.

During the process of checking the data it was necessary to edit some responses and to create missing value categories to identify particular issues such as item non-response. For example, the Household Grid collects the relationships of each household member to the young person, and for parents asks if they are married to or in a cohabiting relationship with anyone else in the household. If the information collected suggested that the parent was a birth parent but in a relationship with a sibling of the young person then this information would be edited. Edits are only carried out if a relevant correction is easily identified (for example, if we know that the household member said to be in a relationship with a parent is not in fact a second parent). If we were unable to identify a correction using the data available (for example, if particular questions should have been asked of the edited cases because we know there is no second parent present, but were not), then a missing value is created. Implausible responses to questions

[2] Details of question wording and routing can be seen in the questionnaires contained in the appendices of the technical report which has also been deposited alongside this user guide.

outside the household grid (e.g. reports by young people of drinking lethal quantities of alcohol) have not generally been edited, but have been noted in an accompanying file (see section 5).

Although no additional cleaning has been applied to the NPD variables during their linkage to this data, all the data linked will have been thoroughly checked and cleaned prior to incorporation into the NPD.

2.8. Datasets

For the purposes of archiving the data it was necessary to remove a number of variables which might compromise the anonymity of the young person and their families. This relates to variables that are either particularly sensitive, or variables where only a small number of respondents answered a particular code. A number of additional variables have been added to replace the most valuable of those that it was necessary to remove and in general these contain codes that have been collapsed to reduce the risk of identifiability through broader categories.

Both the safeguarded and controlled remote access survey datasets cover the following topic areas:

- Household structure derived variables
- Main parent - attitudes to young person's school and involvement in education
- Main parent - extra curricular classes
- Main parent - year 10 subject choices
- Main parent - expectations and aspirations
- Main parent – family activities
- Main parent – relationship with young person
- Main parent – contact with services
- Main parent – reasons for not living with natural parents
- Main parent – risky behaviours
- Main parent – household resources
- Main and second parent – qualifications and education
- Main and second parent – current activity
- Main and second parent – employment/activity history
- Main and second parent – health and demographics
- History parent – birth and health
- History parent – school history
- History parent – choice of current school
- Young person – demographics
- Young person – attitudes to current school
- Young person – year 10 subject choices
- Young person – rules and discipline
- Young person – study support
- Young person – future plans and advice
- Young person – attitudes to school
- Young person – homework
- Young person – relations with parents
- Young person – risk factors
- Young person – household responsibilities
- Young person – employment
- Young person – use of leisure time
- Summary geodemographic data

The controlled remote access datasets also cover the following topics:

- Detailed characteristics

        - Additional language information
        - Detailed age data
        - Detailed geodemographic data
- Income
    - Income variable
- Health
    - Type(s) of medical condition the YP has
- Care
    - Whether living in an institution and the type of institution
    - If has been in care, the type of care and when most recently in care
    - A full set of unedited MP and household structure derived variables (edited in other files to ensure cases in care are not identifiable)
    - Additional YP parental relationship variables (how well YP gets on with their parents, how often they talk to them, whether their parents let them make their own decisions and how often they eat a meal with their family – these are omitted from other files to ensure cases in care are not identifiable)
    - Whether, number and length of periods of MP living apart from the YP

All files contain suitable weights and core sampling variables. At least one school identifier (an anonymous primary sampling unit code) is present on all files, to allow multilevel analysis, with non-anonymised identifiers also available in controlled access datasets.

The content of the various NPD datasets and further detail on school identifiers are discussed in section 4.

Some survey variables are too sensitive to release even in controlled access datasets, such as the household grid.

3. Specifying the sample design

LSYPE2 has a complex sample design incorporating stratification, clustering and unequal selection probabilities within clusters. Standard procedures in SPSS assume a simple random sample (SRS) design. Usually, assuming SRS will tend to underestimate the standard errors. This is because clustering increases variance compared to a SRS design. Hence there is a risk that something statistically insignificant could appear significant. In the case of LSYPE2 the opposite is also possible. This is because the standard weighting procedure in SPSS treats the weighted number of cases as the actual number of cases.

In LSYPE2, the sample was designed to ensure minimum expected wave 7 sample sizes for particular sub-groups of the population including:

- those eligible for free school meals at some point over the preceding three years;
- those eligible for free school meals and have special educational needs of any type;
- each of eight ethnic groups: 'White British', 'Indian', 'Pakistani', 'Bangladeshi', 'Black Caribbean', 'Black African', 'Mixed' and 'Other single ethnic group' and;
- those who attended school in the indepependent sector.

As such, the young people in some of these groups have weights less than 1. So for example, the unweighted number of African (self designated) in the wave 1 sample is 610, the weighted number is 392. Standard procedures in SPSS would calculate the standard error, or a chi square statistic assuming there were 392 cases instead of the actual 610.

If the sample design and the weighting are not properly accounted for, there is a risk of making either a type I or a type II error.

Best practise in order to calculate robust standard errors for LSYPE2 estimates would be for users to specify the sample design in a statistical package such as Stata or SPSS. For example, in Stata users would need to state what the primary sampling units (PSUs) are, what the strata are and what the selection weights are. An alternative option would be the complex samples module in SPSS. In order to use the complex samples options, you would need to specify a file plan, which tells SPSS what the PSUs are, what the strata are and what the selection weights are. However, for this release we have taken the decision not to release the PSUs and strata data for all cases, to avoid identifying the schools attended by those not consenting to NPD linkage. As such, the approach described above will not be possible with the available data.

If you are running cross tabulations and your conclusions are highly significant, e.g. $p < 0.005$, it is almost certain that the conclusion will be significant at the 95% level and you don't need to worry. However, if your conclusion is only just significant at the 95% level, it is advisable to be cautious in your interpretation.

4. Linked Data

Responses to the LSYPE2 survey given by individuals interviewed have been linked to the National Pupil Database (NPD). The NPD contains detailed information about pupils in schools and colleges in England, such as whether young people have free school meals (FSM) or special educational needs (SEN). This database is controlled by the Department for Education, with permission to use data items only granted subject to appropriate data protection. Where both the young person and their parent have given consent for us to do so (both parties consented to the data linking for 93 per cent of the unweighted sample), where matching is possible and where doing so would not automatically risk revealing particularly sensitive characteristics, certain information from the NPD has been matched to their survey responses for use in statistical analysis, further research and report writing.

Further information about the NPD can be found at:

https://www.gov.uk/government/collections/national-pupil-database

Datasets

In order to help users navigate this additional data and, importantly, to protect respondent confidentitality, we have grouped this additional data into a number of datasets. These are available at two levels of security (controlled remote access and controlled saferoom access) dependent on the sensitivity of the information as outlined below.

Controlled remote access data

Controlled data are data which may be identifiable and thus potentially disclosive. These data are only available to users who have been accredited and their data usage has been approved by the relevant Data Access Committee. The user may also be required to undertake specific training as part of such access arrangements. Controlled data we have deposited includes:

1. Data about the young person's KS1 attainment, from 2006. Together with some standard items (surveyID, weights, sample school type flags and primary sampling unit codes), this dataset includes

    a. Type of establishment
    b. National curriculum levels for speaking and listening, reading and writing
    c. National curriculum level for maths
    d. National curriculum levels for science and each contributing topic

2. School census data about the school the young person attended, from 2006, 2010 and 2013 (W1 school). This also includes Ofsted ratings and geodemographic data. In addition to the standard items, this dataset contains

    a. School type
    b. School phase (W1 and 2010 only)
    c. Gender of entry
    d. Admissions policy (W1 only)
    e. Pupil numbers
    f. Percentage with FSM
    g. Percentages with SEN
    h. Percentages with English being or not being their first language
    i. Percentages of pupils in particular ethnic groups
    j. Levels of deprivation based on school postcode (W1 only)
    k. Ofsted rating
    l. Region and local area (W1 only)

m. Urban/rural indicator (W1 only)
n. Banded levels of deprivation based on school postcode, as deciles (W1 only)
o. Anonymised school identifiers (2006 and 2010 only)
p. School identifiers
q. Other detailed geographic variables (W1 only)
r. Levels of deprivation based on school postcode, as raw scores and national ranks (W1 only)
s. Identifiers and school type as at 01/05/13 (W1 only)

3. Data about the KS1 and KS4 levels of attainment in the school the young person attended, from 2006 and 2013 (W1 school) respectively. In addition to the standard items, this dataset contains

a. Percentages achieving various academic thresholds
b. Institution type
c. Number of pupils in the institution (KS4 only)
d. Percentages with SEN (KS4 only)
e. Percentages of the cohort by age (KS4 only)
f. Average point scores (KS4 only)
g. Percentages with FSM (KS4 only)
h. Levels of prior attainment (KS4 only)
i. Percentages with English as an additional language (KS4 only)
j. Percentages achieving expected progress (KS4 only)
k. Time series of level 2 attainment including English and maths (KS4 only)
l. Attainment and entries by prior attainment band (KS4 only)
m. 'Best 8' value added (KS4 only)
n. Numbers of entries (KS4 only)
o. Average point scores per entry (KS4 only)
p. Average grades per qualification (KS4 only)
q. Three-year percentages achieving academic thresholds and progress (KS4 only)

Controlled safe room access data

4. School census data about the young person, from 2006, 2010 and 2013, which is only available through the Data Service's safe room. Together with the standard items, this dataset includes

a. Region
b. Quartiles for levels of deprivation (IDACI) based on pupil postcode
c. School identifiers
d. Enrolment statuses and entry dates
e. Ethnicity codes
f. FSM statuses
g. SEN statuses
h. Whether English is the young person's first language
i. An indicator for changes of school (2010 and 2013 only)
j. Indicators for distances from schools and alternative schools (2013 only)
k. LSOA/LLSOA (ONS lower-level geography)
l. IDACI scores and national ranks
m. Year group
n. Home LA (2010 only)
o. Gifted and talented indicator (2010 only)
p. Mode of travel (2010 only)

Variables

The NPD contains a much wider variety of data than it has been possible to include in this release, such as absence and exclusions histories or census and results data for other years. The variables selected for this release have been prioritised as key indicators of the characteristics and performance of the young person and their school. (Data items of extreme sensitivity, such as those identifying looked after children, were not considered for release.) This release focuses on census and results data from 2006 and 2010 because these were the years when participants will have completed KS1 and KS2, and 2013 as the year participants took part in wave 1.

Where possible, identical variables have been included across all three years, for comparability (e.g. the FSMELIGIBLE flag on the pupil census data). Where this is not possible, any close equivalents have been included (e.g. FIRSTLANGUAGE from 2006 and LANGUAGEGROUPMINOR from 2010 and 2013). The NPD documentation contains detailed definitions of each variable and the years it was in production.

The Ofsted rating system changed from a seven-point to a four-point scale between 2004/05 and 2005/06, rendering the ratings uncomparable – the two versions are reported separately in the 2006 school-based census data (two codes are merged from the larger scale) and careful consideration should be given before any attempt to combine them. (It is also worth noting that the criteria for different Ofsted grades has changed throughout the period, so standards may not be comparable more generally between years.) Ofsted inspection grades have not been assigned to all maintained schools. This is generally because the school will have changed in some way that involves a change of LAESTAB code and there has been no inspection since that change, most commonly due to the creation of a new school or becoming a sponsored academy.

Multiple versions of IDACI are present across the various datasets, relating to different releases of IDACI in different years. (For example, the 2013 school-based census IDACI and IMD come from the 2010 version, as does the non-2007 version in the pupil census.) IDACI may not be consistent between the pupil census version and the version derived from sampled address in the survey data, e.g. because the pupil may have moved between the census and the interview.

Identifying schools and pupils

Types of identifier

There are three main types of school identifier included in the datasets, primary sampling unit codes (PSUs), LAESTAB codes and Unique Reference Numbers (URNs). PSUs have been created as part of the sampling process and have been included to facilitate multilevel analysis, while the latter codes are in general use throughout education data. (LAESTAB codes have 7 digits and are created by concatenating a 3 digit Local Authority identifier with a 4 digit identifier for the establishment, with different ranges for different types of establishment; URNs have 6 digits and do not provide any information beyond serving as an identifier.)

Three different primary sampling units are present, with those included on different datasets depending upon the criteria for access. For the safeguarded data there is only an anonymised version (PSU_A_NPD); controlled access files also have the original code produced during sampling (PSU). (This is built up based on sampling strata, so it can be interpreted to reveal information about the school.) These identifiers are only present for young people that have been matched to the NPD, to prevent the schools attended by non-consenters from being identified. The only dataset with a PSU for all cases is the final survey dataset (the PSU dataset), which contains a separate, anonymous primary sampling unit code (PSU_A) instead of the other two, to allow multilevel analysis of the whole sample. This file is based on the survey controlled dataset , but in order to prevent the indirect release of any information about those not consenting to NPD linkage this file excludes all data that is NPD-derived. This file is only available from DfE upon request; each request will be considered on a case by case basis.

LAESTABs and URNs are held in NPD datasets to identify the schools attended by the young person in the source data. (For example, a particpant's 2006 LAESTAB from the pupil census file will be that of the

school they attended during the 2006 Spring census). They are also shown for the W1 sampled school. Historic LAESTABs and URNs are held in the controlled NPD datasets with additional conditions for reasons including potential sensitivities about the identification of particular institutions or an individual's institution history.

Differences between schools

There are six different schools identified and reported on in the datasets:

- The schools attended by participants during the 2006 and 2010 spring censuses. Data about these schools and the participants while in them is held in the pupil and school-based census datasets, and the schools are identified by the identifiers alongside the 2006 and 2010 census data
- The schools attended by participants during the 2013 spring census. Data about these schools and the participants while in them is held in the pupil census datasets, and the schools are identified by the identifiers alongside the 2013 pupil census data
- The schools where participants sat their KS1 examinations. Data about these schools is held in the pupil and school-based key stage datasets and the the schools are identified by the identifiers alongside the KS1 data.
- The schools participants were considered to be attending at the time of the wave 1 interview. This is the sampled school, unless the participant indicated otherwise during the interview, in which case alternative details they supplied were used to determine the school instead. (If these were insufficient, the 2013 summer census was used instead, unless clearly contradicted by the details supplied.) Data about this school was then drawn from the 2013 spring census. Data about these schools is held in the school-based census and key stage datasets, and the schools are identified by the identifiers alongside the 2013 and KS4 data. The latest Ofsted grade also relates to the W1 school, as at 1st of May 2015.
- The sampled school. This ultimately comes from the autumn 2012 school census. It is only included to allow further consideration of the sampling – sampling data is held in the controlled survey dataset, with identifiers alongside.

Unsurprisingly, the school identifiers will vary between files for many individuals, even within the same year's data – for example, an individual may have moved between the 2006 spring census and that summer's KS1 exams. A further source of change is the move towards academy status - if a school becomes a converter academy the URN changes, but the LAESTAB stays the same; if a school becomes a sponsored academy both its LAESTAB and URN codes change. For this reason, particularly among the 2013/W1/sampling identifiers, differences do not necessarily mean the young person has moved school. (This also complicates the reporting of wave 1 school type – the various school type variables were each true at the cut-off date for the collection [1st of January 2013 for the spring census, 12th of September for the KS4 data, October 2012 for the sampling] , but may have since been superceded by academisation. The wave 1 research report used a reference date of the 1st of May 2013, except for post October-2012 sponsored academies; additional school type and identifier variables as of that date have been included in the controlled school-based census dataset, for anyone wishing to work on a consistent basis.)

Independent schools and PRUs were only required to submit information against seperate, less detailed censuses than traditional maintained mainstream schools. (For example, independent schools are not required to submit any data on individual pupils.) This means that far less information is available about these schools, though it has been included where available; independent and PRU pupils will have missing data for a large majority of NPD variables.

Inclusion of pupils

The NPD files only include pupils for whom we hold full consent for matching, where matching was possible and where doing so did not automatically risk revealing sensitive information about them.

Any off-roll pupils have been removed from the matched data, to reflect that the individual was not part of that institution at the time of the collection. A small number of pupils were interviewed in September

2013 (i.e. in the next academic year), but data for these pupils still relates to the same time period as the rest of the cohort. (Interviewers were instructed to ensure the year 9 school was recorded rather than that for year 10, although there may have been some errors, and to ensure all answers related to the year 9 school.)

5. Questionnaire and Data Problems

It is common in social surveys for issues to arise which can affect data quality. For example, when an interview has to be termininated early so data are only partially collected; when routing errors occur in the questionnaire which mean respondents are not asked question(s) that they should have been asked (or vice versa); interviewer errors (e.g. in coding which person in the household has answered a certain question or set of questions); or even respondent errors (e.g. numeric typos in self-completion sections or questions such as income).

It is therefore worth being mindful of these when analysing LSYPE2 data, as they are potential sources of uncertainty, both in the estimates in the data and the inferences about the results.

As a result, we have released, alongside the data, an excel file *(Known missing or unrealistic data – LSYPE2 W1 release.xls)* which details some of the problems reported during the wave one fieldwork and data processing and also includes a column to identify extreme, unrealistic or missing values.

Missing Data

There can be missing data for a number of reasons. For example, for any survey there can be information missing due to item non-response (a respondent refusing to answer individual questions or not knowing the answers) or errors in survey administration. It is important to consider carefully how to handle missing data in any analysis.

For LSYPE2 there is also missing information among the matched administrative data from the National Pupil Database (NPD), on characteristics such as FSM and SEN status or prior attainment. Where such data is missing from the LSYPE2 dataset, this is either because the respondents have not given consent for the data to be matched, because we did not have sufficient information to match them, because they could not be matched without risking the release of particularly sensitive characteristics, or because the data itself is missing from the NPD.

If we did not receive explicit consent to match to the NPD from both the parent and young person then we have not linked any information from the NPD to the survey responses. For the first wave of LSYPE2, in 7 per cent of cases (unweighted) we did not get explicit consent from both parties. A further 4 per cent (unweighted) gave consent, but we did not hold sufficient information to enable the survey data to be matched to the NPD, most commonly because the young person was attending an independent school, or were not able to do so safely. This left approximately 11,700 young people attending maintained schools for whom we have been able to match in NPD data about them and their school, where available.

There is one further complication: the LSYPE2 cohort sat their KS2 tests in a year when a large number of schools boycotted them. Of those for whom we are able to match in information from the NPD, KS2 test results are missing for slightly less than 30 per cent of cases (unweighted). This also includes around 2 per cent (unweighted) with no KS2 record, which can be the case for a variety of reasons, such as being educated in an independent school or outside England at the time.

We decided to draw the LSYPE2 sample from a complete cohort of pupils, i.e. not to exclude pupils who experienced the KS2 boycott from the sample, because of the substantial risk of unobserved sample bias. Instead, we propose to impute data where possible for those pupils with missing KS2 test results. A

programme of work is planned to undertake this imputation and imputed data will be made available for public use once produced. This programme is also intended to encompass other variables with substantial amounts of missing or implausible data, such as household income.

Given that a substantial and likely unrepresentative minority of the KS2 results are missing, no KS2 results have been included in this first data deposit, either for the young people or their 2010 school.