KANTAR PUBLIC=

LSYPE2 User guide to the datasets: waves 2 & 3

March 2018

Contents

1. Introduction

This user guide provides detailed information pertaining to data arising from waves 2 and 3 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, and coverage of, the survey, the fieldwork procedures, sampling, and weighting please refer to the technical report; this is also available to download from the UK Data Service. Please contact team.longitudinal@education.gov.uk if you require any further information.

This user guide was created jointly by Kantar Public and DfE.

2. Survey Content

2.1 Wave 2 and 3 deposited data

The LSYPE2 Wave 2 and 3 datasets were deposited in March 2018.

There are two levels of access for data deposited on the Data Service:

1. **Safeguarded data** – the main file for each wave contains the majority of the variables from the main parent and young person interviews as well as many of the derived variables.

2. **Controlled access -** Data in this area can be accessed either remotely or within the safe room at Essex University. It contains:

- In a change from wave 1, all survey data is now available as a single dataset for each wave. The main file for each wave has similar content to the safeguarded dataset, additionally containing the majority of sensitive derived and survey variables which are excluded from the safeguarded file. It also contains some high level geodemographic data in the wave 2 file.

- An NPD linked data file containing linked pupil-level KS2 results.

- Two datasets to support analysis with missing data for KS2 attainment for pupils who attended boycott schools in 2010. One contains an inverse probability weight for those who had linked KS2 results and the other dataset has imputed values for KS2 attainment among pupils who attended boycott schools.

The NPD and supporting datasets are discussed further in section 3.

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. The same unique serial number has been used for cohort members across all waves so can be used to link data from different waves to create longitudinal datasets. A typical SPSS command to link files is shown below.

Merging datasets example:

```
GET FILE= ‘C:\ LSYPE2 Wave 2 main file safeguarded release.sav’.
Sort cases by surveyID_W1_ADM (A).
SAVE OUTFILE= ‘C:\ LSYPE2 Wave 2 main file safeguarded release.sav’.
GET FILE= ‘C:\ LSYPE2 Wave 3 main file safeguarded release.sav’
Sort cases by surveyID_W1_ADM (A).
SAVE OUTFILE= ‘C:\ LSYPE2 Wave 3 main file safeguarded release.sav’.

GET FILE= ‘C:\ LSYPE2 Wave 2 main file safeguarded release.sav’.
MATCH FILES /FILE=*
/FILE= 'C:\ LSYPE2 Wave 3 main file safeguarded release.sav’
/BY surveyID_W1_ADM.
SAVE OUTFILE= ‘C:\ LSYPE2 Wave 2 and 3 main file safeguarded release.sav’.
EXECUTE.
```

2.3 Multi-coded variables

Multi-coded 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 multi-coded 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. In the NPD and geodemographic variables, missing data (for whatever reason) is primarily represented by -987,

Box 1: Summary of missing values applied to the LSYPE2 survey 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 derived 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 derived variables to represent data that is not available

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

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

-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

-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).

2.5 Variable names

All survey variable names follow the name given to the variable in the 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, or the main parent
- identify derived variables

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 history module or the young person interview. Geodemographic and derived variables follow the same structure, with question name replaced by a description of the content.

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

Multi-coded variable names are made up of the following characters:

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

A typical 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 multi-coded variable, starting from 1 for the first answer and continuing sequentially
Suffix2	indicates the wave – W2= wave 2; W3 = wave 3 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; 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; GEO = geodemographic variable (based on sample address).

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 code-set 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.

[1] Full details of all derived variables are available in the 'LSYPE2 Wave 2 Derived Variable Documentation' and 'LSYPE2 Wave 3 Derived Variable Documentation' which has also been deposited.

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. 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. Similarly, the derived 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,

Value labels have been appended for all variables, to show what each value represents. If no value label is appended for a response, 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, when a particular case should have been asked a question but was not due to a technical issue. Implausible responses to questions have not generally been edited, but have been noted in the tables in section 4 of this document.

Note that due to back-coding it may appear as if the routing has not been correctly followed for a small number of cases. This can occur when a question used in a filter for subsequent questions has undergone back-coding (when open responses to 'other - specify' questions could actually be coded into one of the original pre-coded response options after the interview). No editing has taken place in the dataset to account for this.

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 Coded variables

Most variables that are shown in the questionnaire as being open ended or those that have an option to specify an other answer have been coded by a specialist team following the interviews. For copies of the questionnaires and for more details on the coding process please refer to the technical report deposited alongside this user guide. For these questions that data contains all codes used in the code frames when coding these variables.

2.9 **Weights**

Both the wave 2 and wave 3 datasets include weights that compensate for attrition since wave 1, as well as differential sampling and response probabilities at wave 1 itself. Users should apply these weights to ensure that the wave 2/wave 3 samples are representative of the LSYPE2 cohort minus those that have exited the population through emigration or death.

Each file contains two weights that are identical apart from the scaling. For example, the wave 2 dataset contains LSYPE2_W2weight_scaled and LSYPE2_W2weight_gross. The ‘scaled’ weight is scaled so that the sum of weights equals the respondent sample size; the ‘gross’ weight is scaled so that the sum of weights equals an estimate of the total population size. Either weight can be used for most analyses but the ‘gross’ weight must be used for estimating population totals with particular characteristics.

Further details on how the weights were specified are available in the technical report for waves 2 and 3.

3. Datasets

3.1 **Survey topics**

Both the safeguarded and controlled remote access main survey datasets for wave 2 cover the following topic areas:

- Household structure summary variables
- Main parent - school history and involvement in education
- Main parent – extra-curricular classes
- Main parent - expectations and aspirations
- 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 – parental relationship history
- Main parent – household resources
- Main parent – employment/activity history
- Main parent – current activity
- Main parent – employment and earnings
- Main parent – qualifications and education
- Main parent – second parent current activity
- Main parent – health and demographics
- History section – birth and health
- History section – history of periods living apart from YP
- Young person – demographics
- Young person – subjects being studied and qualification they are leading to
- Young person – reasons for year 10 subject choices
- Young person – ICT
- Young person – study support
- Young person – future plans and advice
- Young person – knowledge of and intentions towards apprenticeships
- Young person – attitudes to school
- Young person – homework
- Young person – household responsibilities
- Young person – risk factors
- Young person – general health over the last few weeks
- Young person – employment
- Young person – use of leisure time
- Summary geodemographic data

Both the safeguarded and controlled remote access main survey datasets for wave 3 cover the following topic areas:

- Household structure summary variables
- Main parent – attitudes to young person’s school and involvement in education
- Main parent – extra-curricular classes
- 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 – current activity
- Main parent – second parent current activity
- Main parent – income and benefits
- Young person – demographics
- Young person – current activities
- Young person – ICT
- Young person – study support
- Young person – future plans and advice
- Young person – attitudes to school
- Young person – household responsibilities
- Young person – risk factors
- Young person – use of leisure time

Variables relating to either the young person or their parents covering the following topics are only included in controlled datasets, due to their potential sensitivity and small frequencies:

- involvement in criminal/offending/rule breaking behaviour (incl. expulsion and exclusion but not truancy)
- whether they have been victimised/bullied or feel unsafe
- their status regards to special educational needs, learning difficult or mental health
- their sexual orientation
- whether they are or have been in care, adopted, fostered or looked after, or in any way involved with social services

3.2 **Identifiers**

Variables that were used for sampling and a school identifier (an anonymous primary sampling unit code) were deposited with the wave 1 datasets. Users will be able to merge these on to the files for waves 2 and 3 using SurveyID_W1_ADM. For further details about specifying the sample design, please refer to the user guide for the wave 1 datasets.

The content of the KS2 datasets and further detail on school identifiers are discussed in section 3.

Some survey variables are too sensitive to release even in controlled access datasets, such as the household grid. Should there be a particular reason for requiring the household grid dataset, this may be available on request from the Department of Education, subject to data security and research ethics controls. Please contact team.longitudinal@education.gov.uk for more information.

4. Linked Data

Responses to the LSYPE2 survey have been linked to the National Pupil Database (NPD) where consent for this linkage was available. 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. Certain information from the NPD has been matched to the young people's survey responses where:

- Both the young person and their parent have given consent for us to do so (both parties consented to the data linking for 96 per cent of the unweighted sample),
- Matching is possible and
- Doing so would not automatically risk revealing particularly sensitive characteristics,

This linked data is provided 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

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 LAESTAB and URN identifiers 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.) LAESTABs and URNs are held in the controlled NPD datasets to identify the schools attended by the young person.

Datasets

There are three new linked datasets included with this release. An NPD KS2 attainment pupil-level dataset and two complementary datasets to help address missingness in the KS2 data due to the 2010 boycotts are available at two levels of security (controlled remote access and controlled saferoom access) dependent on the sensitivity of the information as outlined below. Any other datasets were released alongside Wave1 survey data and are fully described in the user guide accompanying that wave.

Controlled remote access data

Controlled data are those which might contain information that can be used to identify individual survey participants, disclose their details and so potentially breach data protection legislation. These data are only available to users who have been accredited and whose 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 includes:

1. Data about the young person's KS2 attainment, from 2010. Together with some standard items (surveyID, weights, sample school type flags and primary sampling unit codes), this dataset includes Tier 4 attainment variables
2. Two KS2 datasets using the statistical methods of multiple imputation (MI) and inverse probability weighting (IPW) – both of which can be used by researchers to address issues of possible bias due to missing data. Further information on the creation of these dataset is available below, in the user guide included with this release and in the published technical report.

Controlled safe room access data

The safe room access version of the KS2 results additionally contains the actual primary sampling unit code (PSU) for individuals who consented to NPD linking. An anonymised version of the primary sampling unit code was deposited with the wave 1 data and is also available in the KS2 datasets. This may be sufficient for many types of analysis of data for waves 1 to 3.

Missing KS2 Data

The LSYPE2 cohort sat their KS2 tests in a year when a large number of schools boycotted them. To address the boycott issue, Department for Education (DfE) 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, The DfE commissioned RAND Europe, in collaboration with Professor Vignoles at the University of Cambridge and Professor Brunton-Smith at the University of Warwick, to explore and develop a strategy to address this missing data relating to the boycott.

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. During this work, an inverse probability weight (IPW) for the 8,684 out of 11,823 pupils with responses at Wave 1 who had linked KS2 results, and a dataset of imputed values (MI) for KS2 attainment among pupils who attended boycott schools (for the 8,882 pupils who remain in LSYPE2 at Wave 3) were created. These datasets are available as part of this data release. Advice to analysts faced with the issue of missing data and guidance for users of these datasets are presented in the user guide for the dataset. Full technical details of how these data were created are presented in the accompanying technical report, available at:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/578541/20161205_Technical_report_FINAL.pdf

The MI and IPW datasets can be used with the original KS2 data that is included in this release. The accompanying user guide provides further information and guidance about using these datasets for analysis.

5. Questionnaire and Data Problems

It is common in social surveys for issues to arise which can affect data quality. For example, 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; 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 included a table below which details some of the problems reported during the wave two fieldwork and data processing and also includes a column with some additional comments about some specific variables.

Table 1: Summary of data problems at Wave 2

Variable	**Label**	**Data lost due to known errors**	**Additional comments on variable**
All grid variables	See datasets	8 cases do not have any grid data due to a technical issue during the survey	
ParentCk_MP_W2_MP	Whether interview was conducted jointly with both parents/guardians present	39 cases do not have any data at this question due to a technical issue during the survey	
NPDConMP_W2_MP	Whether MP agreed for NPD linkage permission to continue	17 cases do not have any data at this question due to a technical issue during the survey	
ParentCk_W2_MP	Whether interview was conducted jointly with both parents/guardians present for IP section	39 cases do not have any data at this question due to a technical issue during the survey	
Socchek_W2_MP	Whether job title and description from wave 1 are still correct	9 cases do not have any data at this question due to a technical issue during the survey	
Wrk10_W2_MP	Whether MP has any formal responsibility for supervising the work of other employees	9 cases do not have any data at this question due to a technical issue during the survey	
EmpNum2_W2_MP	Number of different jobs MP has had (since YP born/started living with YP)		Note that due to a script error at wave 1 which meant that some second parents were not able to give correct data for this variable at wave 1, this data was recollected at this variable for any wave 1 second parents who were

			now the main parent and had been affected by the error at wave 1.

Table 2: Summary of data problems at Wave 3

Variable	Label	Data lost due to known errors	Additional comments on variable
NumDiffAdd_W3_GRID	Number of different addresses YP has lived in since last interview		There are 38 cases where the address is reported as incorrect at SameAdd_W3_GRID but this variable is 1. This may be respondent error.
InfoN_W3_GRID	Household member completing the household grid		There are 3 cases with a value of -91 at this variable. These are cases where the young person was living independently without a parent or guardian and in these cases they would have completed the household grid themselves.
NumSchools2_W3_MP	Number of different schools YP has attended since wave 2 interview (where the YP is at a new school)		Note that there are many cases where the value given at this is 1, but it would be expected to be more than this where they currently have a different school. Suspect that some respondents have not been including their current school when answering this question.
JobEarn2_W3_YP	How much money YP earns each week through part-time work during school holidays		2 cases have values outside the range allowed in the questionnaire. This is based on interviewers reporting that respondents' wished to give a higher value than was allowed.
Cgangse_W3_YP	Whether YP is a member of a street gang		There are 40 respondents who don't currently know anyone who is a member of a gang, but who have identified that they are themselves a gang member. This was not a routing error, but was unanticipated.

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.